UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK1120 — Statistiske metoder og dataanalyse 2.
Eksamensdag:	Tirsdag 2. juni 2009.
Tid for eksamen:	14.30 – 17.30.
Oppgavesettet er på 6 sider.	
Vedlegg:	Tabeller over normal-, 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.

I et datasett samlet inn i Sachsen i Tyskland i 1876-1885 angis kjønnsfordelingen i 6125 familier med 12 barn. Dataene sammenfattes i følgende tabell:

Antall jenter	0	1	2	3	4	5	6	7	8	9	10	11	12
Antall familier	7	45	181	478	829	1122	1343	1033	670	286	104	24	3

La O_y være antall familier med eksakt y jenter, $y = 0, 1, \ldots, 12$. Vi skal anta at $(O_0, \ldots, O_{12})$ er multinomisk fordelt med 6125 forsøk og sannsynligheter $(\pi_0, \ldots, \pi_{12})$ der π_y er sannsynligheten for y jenter blant de 12 barna i en familie. Generelt gjelder bare betingelsene $0 \leq \pi_y \leq 1$ og $\sum_{y=0}^{12} \pi_y = 1$, men vi skal undersøke om det er mulig å finne en parametrisering av π_y-ene som passer til dataene.

(a) En rimelig modell kunne være at antall jenter i hver tolvbarnsfamilie er binomisk fordelt med 12 forsøk og sannsynlighet θ for jente i hver fødsel, der θ er den *samme* for alle familier. Vi antar også at familiene er uavhengige. Dermed blir de multinomiske sannsynlighetene $\pi_y = \binom{12}{y}\theta^y(1-\pi)^{12-y}$. Dette kalles i det følgende den binomiske modell.

(Fortsettes side 2.)

Forklar hvorfor forventet antall familier med y barn blir lik

$$\mathrm{E}[O_y] = 6125\binom{12}{y}\theta^y(1-\theta)^{12-y}.$$

Forklar også hvorfor $\hat{\theta} = (\sum_{y=0}^{12} yO_y)/(12 * 6125)(= 0.4807)$ er maximum likelihood estimatet for θ under denne antagelsen.

(b) Nullhypotesen om at π_y følger den binomiske modellen kan testes formelt ved hjelp av en Pearsons kjivadrat-observator

$$X^2 = \sum_{y=0}^{12} \frac{(O_y - E_y)^2}{E_y}$$

der $E_y = 6125\binom{12}{y}\hat{\theta}^y(1-\hat{\theta})^{12-y}$. Angi tilnærmet fordeling for X^2 under nullhypotesen.

For det aktuelle datasettet ble $X^2 = 107.9$. Konkluder om denne modellen passer til dataene.

En alternativ modell er den såkalte beta-binomiske modellen der

$$\pi_y = \binom{12}{y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+y)\Gamma(\beta+12-y)}{\Gamma(\alpha+\beta+12)} = \pi_y(\alpha,\beta) \tag{1}$$

for $y = 0, 1, \ldots, 12$. Her er $\Gamma(\alpha)$ gammafunksjonen og $\alpha > 0$ og $\beta > 0$ er to ukjente parametre. (Modellen er en utvidelse av den binomiske modellen som oppstår ved å anta at antall jenter i ulike familier alle er binomisk fordelt med 12 forsøk, men med *ulike* sannsynligheter θ for hver famile trukket fra en betafordelingen med tetthet $f(\theta;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$ for $0 < \theta < 1$. Du skal **ikke** utlede at dette gir modellen (1).)

(c) Hva blir forventet antall tolvbarnsfamilier med y jenter under den betabinomiske modellen uttrykt ved parametrene (α, β).

De estimerte forventede antallene, der maximum likelihood estimatene (MLE) $(\hat{\alpha}, \hat{\beta})$ er satt inn for (α, β), er oppgitt i Tabell 1 på neste side. Sammenlign med observerte O_y samt anslagene under den binomiske modellen i punkt (a) og (b).

Basert på den beta-binomiske modellen blir Pearsons kjikvadrat-observator nå $X^2 = 13.32$. Angi tilnærmet fordeling for X^2 under en nullhypotese om beta-binomisk modell (1) og konkluder om dataene samsvarer godt med denne modellen.

(Fortsettes side 3.)

Tabell 1: Observerte og estimerte verdier for forventede antall tolvbarnsfamilier sammen med bidrag til Pearsons X^2 for y jenter under binomisk og beta-binomisk modell.

	Observert	Binomisk modell		Beta-binomisk modell	
y	O	E	(O-E)^2/E	E	(O-E)^2/E
0	7	2.4	9.15	5.2	0.63
1	45	26.2	13.53	43.7	0.04
2	181	133.3	17.10	177.9	0.05
3	478	411.2	10.87	462.6	0.51
4	829	856.3	0.87	855.6	0.83
5	1122	1268.1	16.83	1185.3	3.38
6	1343	1369.4	0.51	1261.0	5.33
7	1033	1086.4	2.63	1038.0	0.02
8	670	628.5	2.74	656.0	0.30
9	286	258.5	2.91	310.5	1.93
10	104	71.8	14.45	104.5	0.00
11	24	12.1	11.76	22.4	0.11
12	3	0.9	4.59	2.3	0.20
Total	6125	6125.1	107.9	6125.0	13.32

Oppgave 2.

For å estimere forventet antall tolvbarnsfamilier med y jenter under den beta-binomiske modellen i Oppgave 1c trengte vi maximum likelihood estimatene (MLE) for (α, β). I denne oppgaven skal vi se på hvordan man kan gå fram for å finne disse. Antagelsene er som i Oppgave 1c: $(O_0, \ldots, O_{12})$ er multinomisk fordelt med 6125 forsøk og sannsynligheter $\pi_y(\alpha, \beta)$ gitt ved (1).

(a) Vis at likelihooden for dataene under den beta-binomiske modellen er proporsjonal med

$$L(\alpha, \beta) = \prod_{y=0}^{12} \pi_y(\alpha, \beta)^{O_y}.$$

Vis videre at scorefunksjonen blir lik

$$s(\alpha,\beta) = \begin{bmatrix} 6125(\frac{\Gamma'(\alpha+\beta)}{\Gamma(\alpha+\beta)} - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} - \frac{\Gamma'(\alpha+\beta+12)}{\Gamma(\alpha+\beta+12)}) + \sum_{y=0}^{12} O_y \frac{\Gamma'(\alpha+y)}{\Gamma(\alpha+y)} \\ \\ 6125(\frac{\Gamma'(\alpha+\beta)}{\Gamma(\alpha+\beta)} - \frac{\Gamma'(\beta)}{\Gamma(\beta)} - \frac{\Gamma'(\alpha+\beta+12)}{\Gamma(\alpha+\beta+12)}) + \sum_{y=0}^{12} O_y \frac{\Gamma'(\beta+12-y)}{\Gamma(\beta+12-y)} \end{bmatrix}$$

og finn dessuten et uttrykk for den observerte informasjonsmatrisen (notasjonsmessig er det hensiktmessig å benytte "trigammafunksjonen" $\Psi(\alpha) = \frac{\partial^2 \log(\Gamma(\alpha))}{\partial \alpha^2} = \frac{\partial(\Gamma'(\alpha)/\Gamma(\alpha))}{\partial \alpha}$ i utrykket for den observerte informasjonsmatrisen.)

(Fortsettes side 4.)

(b) Diskuter kort hvordan man på bakgrunn av resultatene i punkt (a) kan beregne MLE for (α, β).

MLE ble $(\hat{\alpha}, \hat{\beta}) = (31.88, 34.44)$ og invers av observert informasjonsmatrise evaluert i $(\hat{\alpha}, \hat{\beta})$ er lik

$$\bar{J}(\hat{\alpha}, \hat{\beta})^{-1} = \begin{bmatrix} 15.43 & 16.64 \\ 16.64 & 18.01 \end{bmatrix}$$

Benytt dette til å beregne et tilnærmet 95% konfidensintervall for α. Gi en begrunnelse for intervallet.

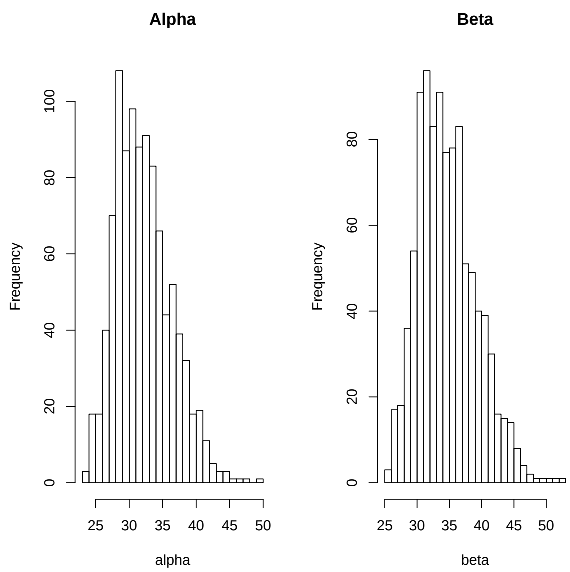
(c) Under er det gjengitt resultater fra en ikke-parametrisk bootstrap med 1000 replikasjoner der $(\alpha_b^\star, \beta_b^\star)$ er MLE for (α, β) under den beta-binomiske modellen for bootstrap-replikasjon b.

Bruk tabellen og figuren nedenfor til å diskutere om tilnærmet inferens basert på MLE $(\hat{\alpha}, \hat{\beta})$ ser ut til å fungere tilfredstillende i dette tilfellet. Momenter til diskusjonen er skjevhet ("bias") og standardfeil for estimatorene, normaltilnærming samt konfidensintervallet fra punkt (b).

Tabell 2: Deskriptiv statistikk for 1000 bootstrap-estimater.

	Gjennomsnitt	Standardavvik	2.5 persentil	97.5 persentil
$\alpha_b^\star$	32.12	4.14	25.20	41.18
$\beta_b^\star$	34.71	4.46	27.29	44.52

Figure 1: Histogram over fordelingen til bootstrap-estimatene $\alpha_b^\star$ og $\beta_b^\star$.



(Fortsettes side 5.)

Oppgave 3.

Modellen for enveis variansanalyse er gitt ved $Y_{ij} = \mu_i + \varepsilon_{ij}$ der μ_i = forventningen i gruppe i og $\varepsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$ er uavhengige, $i = 1, \ldots, I, j = 1, \ldots, J$. Antall replikasjoner i alle grupper er da lik J.
Formålet med oppgaven er å se på sammenhengen mellom den vanlige F-testen for $\mathrm{H}_0 : \mu_1 = \cdots = \mu_I$ og den generaliserte likelihood ratio testen (GLRT) for samme hypotese under den forutsetning at antall grupper I er fiksert, mens antall replikasjoner J vokser. Dermed vil også totalt antall observasjoner $n = IJ$ vokse.

(a) Parameterrommet for modellen gis generelt (i den fulle modellen) ved $\Omega = \{(\mu_1, \ldots, \mu_I, \sigma^2) : \mu_i \in \Re, \sigma^2 \in \Re^+\}$. Beskriv tilsvarende parameterrommet ω_0 under H_0 med en felles forventning μ i alle grupper. Angi dimensjonenene til Ω og ω_0.

Hva er den tilnærmede fordelingen for $-2\log(\Lambda)$ under H_0 (når n er stor) der Λ = generalisert likelihood ratio, dvs. maksimal likelihood under nullhypotesen delt på maksimal likelihood i den fulle modellen.

(b) Angi (uten bevis) eksakt fordeling for F-observatoren $F = MS_B/MS_W$ under nullhypotesen.

Argumenter dessuten for at $(I-1)F$ er tilnærmet kjikvadratfordelt med $I-1$ frihetsgrader under samme nullhypotesen når antall replikasjoner J er stort. Sammenlign med fordelingen for $-2\log(\Lambda)$ i punkt (a).

Her er $MS_W = SS_W/[I(J-1)]$, $SS_W = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \bar{Y}_{i\bullet})^2$, $MS_B = SS_B/(I-1)$, $SS_B = J\sum_{i=1}^{I}(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2$, $\bar{Y}_{i\bullet} = (1/J)\sum_{j=1}^{J} Y_{ij}$, $\bar{Y}_{\bullet\bullet} = (1/n)\sum_{i=1}^{I}\sum_{j=1}^{J} Y_{ij}$. Du kan dessuten bruke at Fisher-fordelingen med ν_1 og ν_2 frihetsgrader defineres som $F_{\nu_1,\nu_2} = (\chi^2_{\nu_1}/\nu_1)/(\chi^2_{\nu_2}/\nu_2)$ der $\chi^2_{\nu_j}$ er to uavhengige kjivadratfordelte stokastiske variable med hhv. ν_1 og ν_2 frihetsgrader.

(c) Log-likelihood for data i en enveis variansanalyse som ovenfor kan skrives

$$l(\mu_1, \ldots, \mu_I, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \mu_i)^2$$

og maksimeres i den fulle modellen for $\hat{\mu}_i = \bar{Y}_{i\bullet}$ (for μ_i-ene) og $\hat{\sigma}^2 = SS_W/n$ (for σ^2) og under nullhypotesen for $\tilde{\mu} = \bar{Y}_{\bullet\bullet}$ (for felles forventning μ) og $\tilde{\sigma}^2 = SS_{TOT}/n$ (for σ^2) der $SS_{TOT} = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \bar{Y}_{\bullet\bullet})^2$. Disse resultatene skal du **ikke** vise. Vis derimot at maksimal log-likelihood $l(\mu_1, \ldots, \mu_I, \sigma^2)$ over Ω kan skrives

$$l(\bar{Y}_{1\bullet}, \ldots, \bar{Y}_{I\bullet}, SS_W/n) = -\frac{n}{2}\log(2\pi\frac{SS_W}{n}) - \frac{n}{2}$$

(Fortsettes side 6.)

samt at maksimal log-likelihood over ω_0 gis ved

$$l(\bar{Y}_{\bullet\bullet}, \ldots, \bar{Y}_{\bullet\bullet}, SS_{TOT}/n) = -\frac{n}{2}\log(2\pi\frac{SS_{TOT}}{n}) - \frac{n}{2}.$$

Vis at vi dermed har

$$-2\log(\Lambda) = n\log(1 + \frac{(I-1)}{I(J-1)}F).$$

Dette betyr at den generaliserde likelihood ratio testen er ekvivalent med den vanlige F-testen. Forklar hvorfor.

(d) Argumenter for at $-2\log(\Lambda) \approx (I-1)F$ når $n = IJ$ er stor og nullhypotesen holder. Sammenhold med fordelingsresulatene i punkt (a) og (b).

SLUTT