UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i	STK2120 — Statistiske metoder og dataanalyse 2
Eksamensdag:	Mandag 6. juni 2011.
Tid for eksamen:	14.30 – 18.30.
Oppgavesettet er på 5 sider.	
Vedlegg:	Tabell over normal, t, χ^2 og F fordeling
Tillatte hjelpemidler:	Godkjent kalkulator og formelsamling for STK1100/STK1110 og STK2120

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

De ulike delpunktene kan stort sett løses uavhengige av hverandre. Hvis du står fast på et punkt, gå derfor heller videre til neste punkt.

Oppgave 1

Tabellen nedenfor viser data fra et eksperiment for å undersøke om avkastning fra en spesifikk kjemisk prosess avhenger enten av formel (Formulat) eller av blandingshastighet (Speed), 3 forsøk for hver kombinasjon.

	Speed		
	60	70	80
Formulat 1	189.7	185.1	189.0
	188.6	179.4	193.0
	190.1	177.3	191.1
Formulat 2	165.1	161.7	163.3
	165.9	159.8	166.6
	167.6	161.6	170.3

Følgende ANOVA tabell er beregnet på det gitte datasettet

```
               Df  Sum Sq Mean Sq  F value Pr > F
Speed           2  230.81  115.41  19.2702 0.0001792
Formulat        1 2253.44 2253.44 376.2702 1.993e-10
Speed*Formulat  2   18.58    9.29   1.5513 0.2516390
Residuals      12   71.87    5.99
```

(Fortsettes på side 2.)

(a) Skriv ned formelen for den modellen som ligger bak ANOVA tabellen som er beregnet.

Hva er estimatet på variansen som inngår i modellen?

(b) Forklar hvilke hypoteser de ulike F-verdier og P-verdier er relatert til og formuler en konklusjon for hver hypotese.

Hvis du skulle forenkle modellen, hva slags modell ville du da foreslå?

Oppgave 2

Tabellen nedenfor viser antall koleratilfeller per husstand i en indisk landsby, dvs av totalt $n = 223$ husstander har 168 av disse 0 tilfeller av kolera osv.

Antall tilfeller	0	1	2	3	4
Frekvens	168	32	16	6	1
Estimert forventning	151.64	58.48	11.28	1.45	0.14

Gjennomsnittelig antall tilfeller er 0.3856.

(a) Siden dette er telledata, vil den første fordeling en tenker på være Poisson fordelingen. Du får her oppgitt at testobservatoren

$$\chi^2 = \sum_{i=0}^{2} \frac{(O_i - E_i)^2}{E_i} = 21.712$$

der O_i er antall husstander med i tilfeller av kolera for $i = 0, 1$ mens O_2 er antall husstander med 2 eller flere tilfeller. E_i er tilhørende forventede antall (estimerte under en Poisson antagelse, gitt fra nederste rad i tabellen ovenfor).

Bruk dette til å utføre en test på om Poisson fordelingen passer til disse data. Formulér en konklusjon.

Begrunn hvorfor det er rimelig å slå sammen kategoriene med 2 eller flere tilfeller.

En alternativ modell til Poisson modellen er følgende:

$$p(y) = \Pr(Y = y) = \begin{cases} \theta & \text{hvis } y = 0 \\ (1-\theta)\frac{1}{1-e^{-\lambda}}\frac{\lambda^y e^{-\lambda}}{y!} & \text{hvis } y \geq 1 \end{cases} \qquad (*)$$

(b) Hva slags begrensninger må en ha på θ?

Vis at hvis $\theta = e^{-\lambda}$ så svarer dette til Poisson modellen.

(Fortsettes på side 3.)

(c) Vis at likelihood funksjonen kan skrives som

$$L(\theta, \lambda) = \prod_{y=0}^{\infty} p(y)^{N_y}$$

der N_y er antall husstander med y tilfeller av kolera.

Vis deretter at log-likelihood funksjonen for de $n = 223$ husstandene er

$$\begin{aligned} l(\theta, \lambda) =& \text{Konst} + N_0 \log(\theta) + (n - N_0) \log(1 - \theta) - \\ & (n - N_0) \log(e^{\lambda} - 1) + \log(\lambda) \sum_{y=1}^{\infty} y N_y. \end{aligned}$$

Hint: Skriv først om $\frac{1}{1-e^{-\lambda}} = \frac{e^{\lambda}}{e^{\lambda}-1}$.

I det etterfølgende vil vi studere algoritmer for å maksimere likelihooden med hensyn på (θ, λ). Følgende egenskaper vil det være behov for:

$$\begin{aligned} E[N_0] =& n\theta \\ E[\sum_{y=1}^{\infty} y N_y] =& n(1 - \theta) \frac{e^{\lambda}}{(e^{\lambda} - 1)} \lambda \end{aligned}$$

Ingen av disse egenskapene behøver du å vise, men kan brukes i det etterfølgende.

(d) Skriv ned definisjonen av score-funksjonen. Vis at score-funksjonen er gitt ved

$$\boldsymbol{s}(\theta, \lambda) = \begin{pmatrix} \frac{N_0}{\theta} - \frac{n - N_0}{1 - \theta} \\ -(n - N_0) \frac{e^{\lambda}}{e^{\lambda} - 1} + \frac{1}{\lambda} \sum_{y=1}^{\infty} y N_y \end{pmatrix}$$

Regn ut forventningen til $\boldsymbol{s}(\theta, \lambda)$. Hvordan stemmer dette med den generelle teori?

En kan også vise at den observerte informasjonsmatrisen er gitt ved

$$\bar{\boldsymbol{J}}(\theta, \lambda) = n \begin{pmatrix} \frac{N_0}{\theta^2} + \frac{n - N_0}{(1-\theta)^2} & 0 \\ 0 & -(n - N_0) \frac{e^{\lambda}}{(e^{\lambda}-1)^2} + \frac{1}{\lambda^2} \sum_{y=1}^{\infty} y N_y \end{pmatrix}$$

mens den forventede (Fisher) informasjonsmatrisen er lik

$$\bar{\boldsymbol{I}}(\theta, \lambda) = n \begin{pmatrix} \frac{1}{\theta} + \frac{1}{1-\theta} & 0 \\ 0 & (1 - \theta) \frac{e^{\lambda}}{e^{\lambda}-1} \left[\frac{1}{\lambda} - \frac{1}{e^{\lambda}-1} \right] \end{pmatrix}$$

Dette behøver du *ikke* å vise.

(Fortsettes på side 4.)

(e) Diskuter hvilke generelle fordeler scoring algoritmen har i forhold til Newton-Raphson algoritmen.

Hva slags konsekvenser har det mhp beregningene i algoritmene at informasjonsmatrisene er diagonale?

Tabellen nedenfor viser $(\theta^{(s)}, \lambda^{(s)}, l^{(s)})$ for $s = 1, 2, \ldots$ for scoring algoritmen der s er iterasjonssnummer.

Iterasjon	$\theta^{(s)}$	$\lambda^{(s)}$	$l^{(s)}$
0	0.5000	1.0000	-184.3447
1	0.7534	0.9715	-154.3288
2	0.7534	0.9722	-154.3287
3	0.7534	0.9722	-154.3287

(f) Merk spesielt at verdien på θ ikke endrer seg etter første iterasjon. Vis at dette ikke er en tilfeldighet men en egenskap ved algoritmen i dette tilfellet.

(g) Vis at $\hat{\theta}$ kunne blitt funnet direkte ved analytiske beregninger og at dette samsvarer med det du fikk i (f).

Er dette estimatet rimelig?

Hvordan kunne en utnytte dette for å lage en enklere optimeringsprosedyre?

(h) Anta vi ønsker å bruke en normaltilnærming for å si noe om egenskapene til ML estimatene. Hva slags betydning har det at informasjonsmatrisene er diagonale i dette tilfellet?

Vi får i dette tilfellet at

$$\bar{\boldsymbol{I}}(\hat{\theta}, \hat{\lambda}) = \begin{pmatrix} 1200.170 & 0.0000 \\ 0.000 & 37.1737 \end{pmatrix}$$

Bruk dette til å lage 95% konfidensintervall både for θ og λ.

(i) Et alternativ til normaltilnærming er å bruke bootstrapping.

Hvis vi simulerer y_i^* ved å trekke fra modell (*), hva slags bootstrapping prosedyre har vi da?

Nedenfor er ulike oppsummerende mål av bootstrapsimuleringene for $\hat{\theta}$ og $\hat{\lambda}$ angitt.

	Gj.snitt	Standard avvik	Nedre 2.5% kvantil	Øvre 2.5% kvantil
$\hat{\theta}^*$	0.755	0.029	0.695	0.812
$\hat{\lambda}^*$	0.971	0.167	0.656	1.298

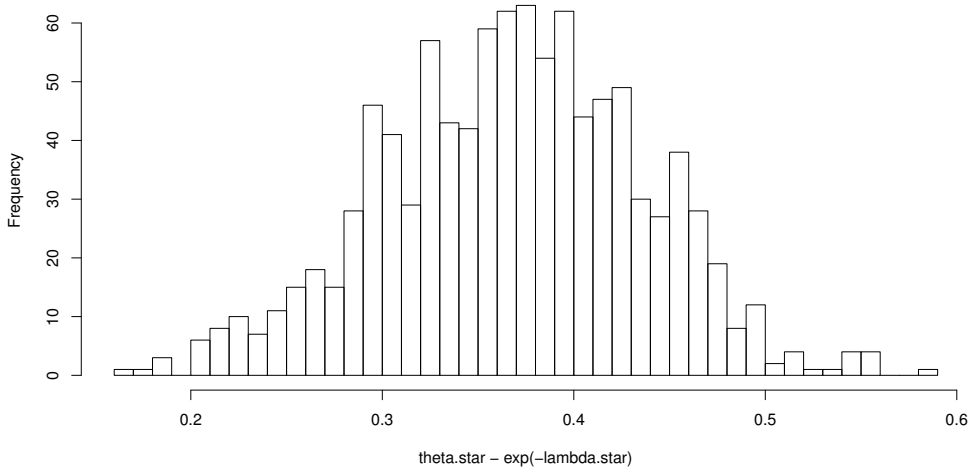
Lag 95% standard bootstrap konfidensintervall for de to parametrene.

Kommenter eventuelle forskjeller fra det du fikk ved normaltilnærmingen.

(Fortsettes på side 5.)

(j) Nedenfor er vist et histogram av $\hat{\theta}^* - e^{-\hat{\lambda}^*}$ basert på bootstrap simuleringene. Diskutér hvordan dette kan brukes til en alternativ test på om dataene følger en Poisson fordeling.

(Du skal her ikke utføre noen test, kun skissere hvordan dette kan gjøres.)



SLUTT