Københavns Universitet
Det Naturvidenskabelige Fakultet

Eksamen i Statistik 1

14. april 2016

Eksamen varer 4 timer. Alle hjælpemidler er tilladt under hele eksamen, men du må ikke have internetforbindelse. Besvarelsen må gerne skrives med blyant.

Eksamenssættet består af fire opgaver med i alt 14 delspørgsmål. Alle delspørgsmål vægtes ens. Data til opgave 3 ligger i filen `hydrolyse` på en USB-stick. Sticken skal afleveres tilbage når eksamen slutter, men udelukkende for at den kan genbruges. Den kan altså ikke indgå som del af besvarelsen.

Opgave 1

Lad $t_1, \ldots, t_n$ være kendte tal, og lad $X_1, \ldots, X_n$ være uafhængige stokastiske variable hvor X_i er normalfordelt med middelværdi $e^{\beta t_i}$ og varians σ^2. Fordelingen af $(X_1, \ldots, X_n)$ afhænger af den ukendte parameter $\theta = (\beta, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+$. Bemærk at der *ikke* er tale om en lineær normal model da middelværdivektoren ikke antages at ligge i et underrum.

1. Opskriv log-likelihoodfunktionen og scorefunktionen for en observation $x = (x_1, \ldots, x_n)$. Bemærk at parameteren er to-dimensional, således at scorefunktionen har værdier i $\mathbb{R}^2$.

   Betragt desuden datasættet bestående af følgende værdier:

   ```
   > t
   [1] 0.0 0.6 1.2 1.8 2.4 3.0
   > x
   [1]  3.0  1.6  2.8  9.2 17.5 31.4
   ```

   Eftervis at $(\beta, \sigma^2) = (1.15625068, 1.606481)$ løser scoreligningerne (på nær afrunding). Skriv fx en stump R-kode der viser de beregninger du har lavet.

Løsningen til scoreligningerne er faktisk et maksimum likelihood estimatet. Nedenfor kan afrunding til tre decimaler benyttes således at MLE for det givne datasæt er $(\hat{\beta}, \hat{\sigma}^2) = (1.156, 1.606)$.

2. Vis at elementerne på plads (1,1) og plads (1,2) i Fisherinformationsmatricen er givet ved

$$i(\beta, \sigma^2)_{1,1} = \frac{1}{\sigma^2} \sum_{i=1}^{n} t_i^2 e^{2\beta t_i}, \quad i(\beta, \sigma^2)_{1,2} = 0.$$

3. Angiv den asymptotiske varians for $\hat{\beta}$ for vilkårlige observationer $(x_1, \ldots, x_n)$. Betragt derefter det givne datasæt fra spørgsmål 1, og bestem et estimat for spredningen af $\hat{\beta}$ samt et (approksimativt) 95% konfidensinterval for β.

   Du kan bruge de sædvanlige resultater vedrørende asymptotisk fordeling af MLE uden bevis.

Antag nu at $x_i > 0$ for alle $i = 1, \ldots, n$ og lad $z_i = \log(x_i)$. Som alternativ model for effekten af t kan vi betragte den lineære normale model hvor vi antager at $z_1, \ldots, z_n$ er udfald af uafhængige stokastiske variable $Z_1, \ldots, Z_n$ med $Z_i \sim N(\gamma t_i, \tau^2)$ hvor $\gamma \in \mathbb{R}$ og $\tau^2 > 0$ er ukendte parametre.

4. Bestem maksimum likelihood estimatet $\hat{\gamma}$, både for vilkårlige observationer $z_1, \ldots, z_n$ og for datasættet fra spørgsmål 1. Bestem desuden den estimerede spredning for estimatoren $\hat{\gamma}$ for det givne datasæt.

Opgave 2

Lad X være normalfordelt på $\mathbb{R}^3$ med middelværdi og varians givet ved

$$\mathrm{E}X = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad \mathrm{V}X = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 9 & 1 \\ 1 & 1 & 4 \end{pmatrix},$$

og definer desuden den stokastiske variabel Y ved

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} X_1 - X_2 + X_3 \\ X_1 + 2X_2 - 3X_3 \end{pmatrix}.$$

1. Find fordelingen af Y, og afgør om Y er regulært eller singulært normalfordelt.
2. Bestem alle par $(c_1, c_3) \in \mathbb{R}^2$ således at fordelingen af $Z = c_1X_1 + c_3X_3$ opfylder følgende to betingelser:
   - Z er normalfordelt med middelværdi 0 og varians 21
   - Z og $X_1 + X_3$ er uafhængige

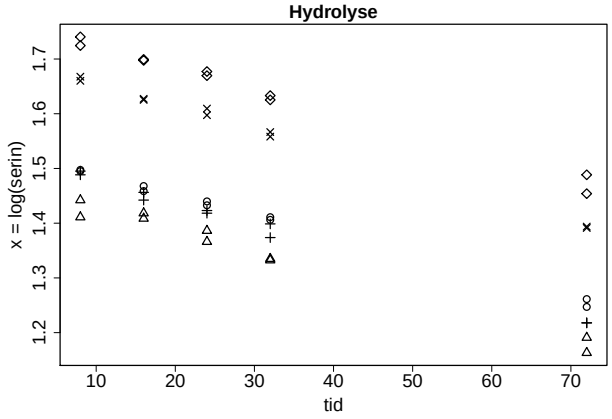
Opgave 3

Hydrolyse er en kemisk reaktion eller proces, hvor et molekyle reagerer med vand og bliver opdelt i mindre molekyler. I et eksperiment har man undersøgt effekten af hydrolysetiden på mængden af aminosyren serin for fem forskellige fodertyper. Fem forskellige hydrolysetider blev testet, og der er to observationer for hver af de 25 kombinationer af fodertype og hydrolysetid. Datasættet består således af 50 observationer.

Data er tilgængelige i filen `hydrolyse.txt` på den vedlagte USB-stick. Der er tre variable:

- `foder`: Faktor der angiver fodertypen. Der er fem labels (mulige værdier): `byg`, `fiskemel`, `majs`, `kb.mel` (kød- og benmel), `soja`.
- `tid`: Hydrolysetid i timer. Har værdierne 8, 16, 24, 32, 72.
- `serin`: Mængden af serin efter hydrolyse i enheden $g/16gN$.

Nedenfor er vist et scatterplot af tid og log-serinmængde, dvs. `log(serin)`. Observationer hørende til forskellige fodertyper er vist med forskellige symboler.



Figuren antyder at det kunne være fornuftigt at lade middelværdien af log-serinmængden for de fem fodertyper være lineære funktioner af hydrolysetiden. Vi skal derfor betragte to modeller:

```
modelA <- lm(log(serin) ~ foder*tid, data=hydrolyse)
modelB <- lm(log(serin) ~ foder+tid, data=hydrolyse)
```

I kommandoerne er `hydrolyse` navnet på et R-datasæt hvor data er indlæst.

1. Angiv den statistiske model der svarer til `modelA` og udfør modelkontrol for modellen. Du skal skitsere og kommentere de figurer du laver.

2. Undersøg med et hypotesetest om det er rimeligt at antage at hældningerne er ens for alle fem fodertyper.

I resten af opgaven skal du kun bruge `modelB`.

3. Bestem et estimat og et 95% konfidensinterval for følgende forskelle:

   - Forskellen mellem forventet log-serinmængde for byg og majs
   - Forskellen mellem forventet log-serinmængde for majs og soja

4. Bestem et estimat og et 95% konfidensinterval for forskellen mellem forventet log-serinmængde ved 20 og 50 timers hydrolysetid (for fastholdt fodertype).

   Bestem også den prædikterede værdi og et 95% prædiktionsinterval for log-serinmængden efter 40 timer for rotter der fodres med soja.

Lad $\gamma_1, \ldots, \gamma_5$ være middelværdierne af log-serin efter 50 timers hydrolysetid for hver af de fem fodertyper. Gennemsnittet $\bar{\gamma} = \frac{1}{5}(\gamma_1 + \gamma_2 + \gamma_3 + \gamma_4 + \gamma_5)$ kan så fortolkes som forventet log-serinmængde for en „gennemsnitsfodertype“ efter 50 timers hydrolyse.

5. Bestem et estimat og et 95% konfidensinterval for $\bar{\gamma}$.

   Vink: Hvordan kan $\bar{\gamma}$ skrives som funktion af parametrene i modellen? Det er lidt nemmere (men ikke nødvendigt) at benytte en anden parametrisering af modellen end den i `modelB`.

Opgave 4

Lad $X_1, \ldots, X_n$ være uafhængige stokastiske variable der alle har tæthed f_θ mht. tællemålet på $\mathbb{N}_0$, hvor

$$f_\theta(x) = (1-\theta)^x \theta, \quad x \in \mathbb{N}_0$$

Fordelingen afhænger af den ukendte parameter $\theta \in (0,1)$.

For en observation $x = (x_1, \ldots, x_n) \in \mathbb{N}_0^n$ er log-likelihoodfunktionen således givet ved

$$\ell_x(\theta) = -\log L_x(\theta) = -n\log(\theta) - \log(1-\theta)\sum_{i=1}^{n} x_i, \quad \theta \in (0,1)$$

Dette kan uden videre benyttes i det følgende.

1. Vis at maksimum likelihood estimatoren for θ er givet ved

$$\hat{\theta} = \frac{n}{n + \sum_{i=1}^{n} X_i}.$$

2. Betragt datasættet bestående af følgende 12 tal:

$$10 \quad 2 \quad 3 \quad 0 \quad 0 \quad 1 \quad 1 \quad 0 \quad 2 \quad 7 \quad 2 \quad 5$$

   Bestem maksimum likelihood estimatet, $\hat{\theta}$, for disse data.

   Betragt derefter hypotesen $H : \theta = 0.4$, og udfør likelihood ratio testet (kvotienttestet) for hypotesen. Du kan bruge det sædvanlige asymptotiske resultat vedrørende $LR(X) = -2\log Q(X)$ uden bevis.

3. Betragt stadig hypotesen $H : \theta = 0.4$. I dette spørgsmål skal du lave et simulationsstudie hvor du undersøger styrken af likelihood ratio testet under forskellige omstændigheder. Mere præcist skal du for forskellige værdier af n og den sande værdi af θ (se tabellen nedenfor) gøre følgende:

   - Simulere et udfald af $x = (x_1, \ldots, x_n)$ vha. kommandoen `rgeom(n,theta)` hvor `n` og `theta` har de relevante værdier.
   - Udføre likelihood ratio testet for hypotesen $H : \theta = 0.4$ for det simulerede x. Du kan bruge det sædvanlige asymptotiske resultat vedr. $LR(X) = -2\log Q(X)$ uden bevis.
   - Gentage dette 5000 gange og registrere med hvilken relativ hyppighed hypotesen forkastes.

   Besvarelse: Besvarelsen af spørgsmålet skal bestå af følgende ting:

   - En udfyldt version af nedenstående skema:

n	θ	Relativ hyppighed hvormed hypotesen forkastes
20	0.4	**
20	0.5	**
20	0.6	**
40	0.4	**
40	0.5	**
40	0.6	**

   - Kommentarer til dine resultater