Københavns Universitet
Det Naturvidenskabelige Fakultet

Eksamen i Statistik 1

9. april 2015

Eksamen varer 4 timer. I de sidste to timer, men ikke de første to timer, er det tilladt at benytte computer uden internetforbindelse. Andre hjælpemidler (dog ikke tablets) er tilladt under hele eksamen. Besvarelsen må gerne skrives med blyant.

Eksamenssættet består af tre opgaver med i alt 13 delspørgsmål. Alle delspørgsmål vægtes ens. Data til opgave 3 ligger i filen `heste-L2.txt` 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

Betragt fordelingen der har tæthed

$$f_\theta(x) = \begin{cases} \theta\left(\frac{1}{2}\right)^{x+1} & \text{for } x \in \mathbb{N} \\ 1 - \frac{\theta}{2} & \text{for } x = 0 \end{cases}$$

mht. tællemålet på $\mathbb{N}_0$. Fordelingen afhænger af parameteren θ.

Du kan uden bevis benytte at f_θ faktisk er en tæthed for alle $\theta \in (0,2)$. Det kan ligeledes benyttes uden bevis at der for $|a| < 1$ gælder

$$\sum_{k=1}^{\infty} k \cdot a^k = \frac{a}{(1-a)^2} \quad \text{og} \quad \sum_{k=1}^{\infty} k^2 \cdot a^k = \frac{a(1+a)}{(1-a)^3}$$

Lad $X_1, \ldots, X_n$ være uafhængige stokastiske variable der alle har tæthed f_θ med ukendt $\theta \in (0,2)$. Definér desuden $N_0 = \sum_{i=1}^n \mathbf{1}_{\{X_i=0\}}$, dvs. den stokastiske variabel der tæller hvor mange af X_i'erne der er lig 0.

1. Definer estimatoren $\tilde{\theta} = \frac{1}{n}\sum_{i=1}^n X_i$. Vis at $\tilde{\theta}$ er central for θ, og bestem desuden $\mathrm{V}_\theta \tilde{\theta}$.

2. Lad $x = (x_1, \ldots, x_n) \in \mathbb{N}_0^n$. Gør rede for at likelihoodfunktionen på nær en proportionalitetsfaktor er givet ved

   $$L_x(\theta) = \left(1 - \frac{\theta}{2}\right)^{n_0} \theta^{n-n_0}, \quad \theta \in (0,2)$$

   hvor $n_0 = \sum_{i=1}^n \mathbf{1}_{\{x_i=0\}}$ er antallet af nuller blandt tallene $x_1, \ldots, x_n$.

   Bestem derefter scorefunktionen og den observerede informationsfunktion.

3. Vis at hvis $N_0 \notin \{0, n\}$, så findes maksimum likelihood estimatoren og er givet ved

   $$\hat{\theta} = \frac{2(n - N_0)}{n}. \tag{1}$$

Det er nemt at se, at hvis $N_0 \in \{0, n\}$, så angiver formel (1) den værdi af θ i det *afsluttede* interval $[0, 2]$ der maksimerer L_X. I det følgende betragter vi $\hat{\theta}$ som veldefineret og givet ved (1) uanset værdien af N_0, altså for alle $N_0 \in \{0, 1, \ldots, n\}$.

4. Angiv fordelingen af N_0. Bestem derefter middelværdi og varians for $\hat{\theta}$. Vil du foretrække $\tilde{\theta}$ eller $\hat{\theta}$ som estimator for θ?

5. Betragt datasættet bestående af følgende 15 tal:

   3 6 2 3 0 1 2 2 1 0 4 1 1 2 0

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

   Betragt derefter hypotesen $H : \theta = 1$. Bestem $\log L_x(\hat{\theta})$ og $\log L_x(1)$ for de givne data, og udfør likelihood ratio testet (kvotienttestet) for hypotesen. Du kan bruge det sædvanlige asymptotiske resultat vedrørende $LR(X) = -2\log Q(X)$ uden bevis.

6. Betragt igen hypotesen $H : \theta = 1$. I dette spørgsmål skal du lave et simulationsstudie hvor du undersøger hvor ofte hypotesen forkastes under forskellige omstændigheder. Bemærk at likelihood ratio teststørrelsen kun afhænger af x via n_0. Derfor kan man nøjes med at simulere værdier af N_0.

   Mere præcist skal du for forskellige værdier af n og den sande værdi af θ (se tabellen nedenfor) gøre følgende:

   - Simulere et udfald af N_0.
   - Udføre likelihood ratio testet for hypotesen $H : \theta = 1$ for det simulerede udfald. Du kan bruge det sædvanlige asymptotiske resultat vedr. $-2\log Q(X)$ uden bevis.
   - Gentage dette 5000 gange og registrere med hvilken relativ hyppighed hypotesen forkastes.

   Vink til simulationsdelen: Du kan bruge funktionen `rbinom` til simulation af N_0. For eksempel giver kommandoen `rbinom(n=1, size=5, prob=0.25)` et udfald fra binomialfordelingen med antalsparameter 5 og sandsynlighedsparameter 0.25.

   Besvarelse: Besvarelsen af spørgsmålet består af følgende ting:

   - En udfyldt version af nedenstående skema:

n	θ	Relativ hyppighed hvormed hypotesen forkastes
50	1	**
250	1	**
50	1.2	**
250	1.2	**
50	1.4	**

   - Kommentarer til dine resultater

Opgave 2

Lad X_1, X_2 og X_3 være reelle stokastiske variable med følgende egenskaber:

- $\mathrm{E}X_i = 0$ for alle $i = 1,2,3$
- X_3 og $(X_1,X_2)^T$ er uafhængige
- $(X_1,X_2)^T$ er normalfordelt på $\mathbb{R}^2$ med variansmatrix $\begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$
- $X_3 \sim N(0,9)$

Lad endelig $X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$ og definér desuden $Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} X_1 + X_2 + X_3 \\ X_1 + X_2 - X_3 \end{pmatrix}$

1. Opskriv fordelingen af X, og afgør om X er regulært normalfordelt.
2. Bestem fordelingen af Y, og afgør om Y er regulært normalfordelt. Er Y_1 og Y_2 uafhængige?
3. Bestem $c \in \mathbb{R}$ således at $Z = X_1 + c \cdot X_2$ er konstant med sandsynlighed 1.

Opgave 3

Halthedsundersøgelse af en hest består blandt andet af visuel vurdering af hesten når den løber. Denne vurdering er subjektiv, og der kan være stor forskel på forskellige dyrlægers vurdering af den samme hest. En gruppe dyrlæger på Københavns Universitet arbejder derfor med at kvantificere graden af halthed på en objektiv måde. De benytter kontinuerte målinger af hestens accelerationer mens den traver.

Denne opgave handler om et halthedsmål der betegnes $L2$. Værdien af $L2$ er per konstruktion strengt positiv, og udfra definitionen vil man forvente små værdier for raske heste og store værdier for halte heste. Formålet med denne opgave er blandt andet at undersøge om fordelingen af $L2$ faktisk er forskellig for raske og halte heste.

Data stammer fra et forsøg med 85 heste. Hver hest er enten rask (ikke halt) eller halt på et af de fire ben, og for hver hest har man målt $L2$. Data er tilgængelige i filen `heste-L2.txt` og består af følgende variable, der er registreret for hver af de 85 heste:

- `grp`: Faktor der angiver om hesten er rask (værdien `Rask`), eller hvilket ben hesten er halt på (`HF` for højre forben, `HB` for højre bagben, `VF` for venstre forben, og `VB` for venstre bagben)
- `L2`: Numerisk variabel med værdien af $L2$-størrelsen

Det er en god ide, men ikke et krav, at lave en eller flere figurer af data før analysen.

Strukturen af datasættet lægger op til en ensidet variansanalyse med `grp` som faktor. Hvis responsvariablen betegnes $X = (X_1, \ldots, X_{85})^T$, er antagelsen altså at $X \sim N(\xi, \sigma^2 I)$, hvor $\xi \in L_{\texttt{grp}}$ og $\sigma^2 > 0$ er ukendte parametre.

I spørgsmål 1 skal du benytte denne model med to forskellige responsvariable: `L2` henholdsvis `log(L2)`.

1. Fit begge modeller i R, dvs. en model for hver responsvariabel, og lav residualplot for hver af dem. Gør kortfattet rede for hvilken af modellerne der er mest egnet til at beskrive variationen i data.

   Besvarelsen skal indeholde:

   - De to kommandoer du har brugt til at fitte modellerne
   - Skitser af de to residualplot
   - En kort redegørelsen for hvilken model der er mest egnet til at beskrive variationen i data.

I resten af opgaven skal du kun benytte den model som du fandt mest egnet. Fra nu af er X altså enten `L2` eller `log(L2)`, afhængig af din konklusion i spørgsmål 1. Lad desuden α_{Rask}, α_{HF}, α_{HB}, α_{VF} og α_{VB} betegne middelværdien af X_i i de fem grupper.

2. Angiv et estimat for α_{Rask}. Angiv også fordelingen af estimatoren $\hat{\alpha}_{\text{Rask}}$ samt den estimerede spredning i denne fordeling, dvs. standard error for $\hat{\alpha}_{\text{Rask}}$.

3. Opskriv og test hypotesen om at fordelingen af X_i er den samme for alle fem grupper.

   Undersøg derefter, med et enkelt hypotesetest, om fordelingen af X_i er den samme for de fire grupper hvor hestene er halte (`HF`, `HB`, `VF`, `VB`).

4. Angiv et estimat og et 95% konfidensinterval for forskellen mellem den forventede værdi af X_i for halte og raske heste. Er fordelingen af X_i den samme for halte og raske heste?