Københavns Universitet
Det Naturvidenskabelige Fakultet

Eksamen i Statistik 1

29. juni 2017

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 tre opgaver med i alt 13 delspørgsmål. Alle delspørgsmål vægtes ens. Data til opgave 3 ligger i filen `lim.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

Lad $t_1,\ldots,t_n$ være kendte, positive tal, og lad $X_1,\ldots,X_n$ være uafhængige stokastiske variable hvor X_i har tæthed

$$f(x_i) = \theta t_i \cdot x_i^{\theta t_i - 1} \cdot 1_{(0,1)}(x_i)$$

mht. lebesguemålet på $\mathbb{R}$. Her er $\theta > 0$ en ukendt parameter. De stokastiske variable er altså uafhængige, men ikke identisk fordelte.

Du kan benytte de sædvanlige asymptotiske resultater vedr. fordelingen af maksimum likelihood estimatoren og kvotienttesstørrelsen uden bevis (de gælder selvom X_i'erne ikke er identisk fordelte).

1. Opskriv likelihoodfunktionen for en observation $x = (x_1,\ldots,x_n) \in (0,1)^n$, og vis at scorefunktionen er givet ved
$$S_x(\theta) = -\frac{n}{\theta} - \sum_{i=1}^{n} t_i \log(x_i).$$
Bestem desuden Fisherinformationen.

2. Bestem maksimum likelihood estimatoren $\hat{\theta}$ for θ, og angiv dens asymptotiske fordeling.

3. Definér $Y_i = -t_i \log(X_i)$ og $S_Y = \sum_{i=1}^{n} Y_i$. Vis at θS_Y er gammafordelt med formparameter n og skalaparameter 1, og bestem et eksakt 95% konfidensinterval for θ.

Betragt datasættet bestående af følgende observationer ($n = 10$):

```
> t
 [1] 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
> x
 [1] 0.88 0.86 0.62 0.79 0.66 0.54 0.88 0.87 0.73 0.79
```

4. Vis at likelihood ratio teststørrelsen er givet ved
$$LR(\theta,x) = -2\log Q(\theta,x) = 2n\left(-\log\theta + \theta\bar{y} - \log\bar{y} - 1\right)$$
hvor $\bar{y} = -\frac{1}{n}\sum t_i \log x_i$. Benyt resultatet til at teste hypotesen $H : \theta = 4$ for det givne datasæt.

5. Bestem $\hat{\theta}$ og det eksakte 95% konfidensinterval fra spørgsmål **??** for det givne datasæt. Gør desuden rede for at 95% konfidensintervallet baseret på den approksimative pivot $LR(\theta, X) = -2\log Q(\theta, X)$ for det givne datasæt er $(3.1754, 11.1151)$.

6. I dette spørgsmål skal du ved hjælp af simulation undersøge middelværdi og spredning af maksimum likelihood estimatoren $\hat{\theta}$. Mere præcist skal du

   - Simulere data fra modellen for $n = 10$, $n = 25$ og $n = 250$. Den sande værdi af θ er hele tiden 5, og $t_i = i/n$. For $n = 10$ kan dette gøres ved hjælp af kommandoerne

     ```
     t <- (1:10)/10
     x <- rbeta(10, shape1=5*t, shape2=1)
     ```

   - Beregne $\hat{\theta}$
   - Gentage dette 5000 gange, beregne gennemsnit og empirisk spredning af estimaterne og sammenligne med den asympotiske fordeling

   Besvarelse: Besvarelsen af spørgsmålet består af følgende ting:

   - En udfyldt version af nedenstående skema:

		Simulation		Asymptotisk fordeling	
n	θ	gennemsnit	spredning	middelværdi	spredning
10	5	**	**	**	**
25	5	**	**	**	**
250	5	**	**	**	**

   - Kommentarer til dine resultater, herunder: Er $\hat{\theta}$ en central estimator for θ?

Opgave 2

Lad X være normalfordelt på $\mathbb{R}^3$ med middelværdi 0 og varians Σ, altså $X \sim N(0, \Sigma)$, hvor

$$\Sigma = \begin{pmatrix} 5 & -1 & 6 \\ -1 & 6 & -7 \\ 6 & -7 & 13 \end{pmatrix}.$$

1. Bestem fordelingen af $Y = \begin{pmatrix} X_1 - X_2 \\ X_2 + 2X_3 \end{pmatrix}$. Er Y regulært eller singulært normalfordelt?

2. Gør rede for at X er singulært normalfordelt på $\mathbb{R}^3$, men regulært normalfordelt på mængden $\{x \in \mathbb{R}^2 \,|\, x_3 = x_1 - x_2\}$.

Opgave 3

Et firma der fremstiller trælim, har tre virksomme komponenter til rådighed. Komponenterne kaldes A, B og C. Komponenterne kan indgå alene eller i par, men alle tre kan ikke tilsættes samtidig. Der er således syv forskellige muligheder:

$$\text{Ingen komponenter},\ A,\ B,\ C,\ AB,\ AC,\ BC$$

hvor fx A betyder at kun komponent A er tilsat, mens fx AB betyder at komponent A og B begge er tilsat.

For at teste effekten af de tre komponenter, har firmaet udført et eksperiment hvor de syv muligheder er afprøvet otte gange hver. Hver gang blev to træklodser limet sammen, og styrken af limningen blev målt efter en times tørretid. Høje værdier indikerer stor styrke.

Data er tilgængelige i filen `lim.txt` på den vedlagte USB-stick. Der er 56 datalinier og fire variable:

- `styrke`: Styrke af limning
- `A`, `B`, `C`: Numeriske variable med værdien 1 hvis den pågældende komponent er benyttet, og 0 ellers

I det følgende er $Y_1, \ldots, Y_{56}$ de stokastiske variable hørende til de 56 styrkemålinger. Det antages overalt at $Y_1, \ldots, Y_n$ er uafhængige og normalfordelte variable med samme varians. I de første fire spørgsmål skal du desuden antage at middelværdierne har følgende form:

$$\mathrm{E}Y_i = \alpha + \beta_1 A_i + \beta_2 B_i + \beta_3 C_i, \quad i = 1, \ldots, 56$$

hvor A_i, B_i og C_i er 0 eller 1 alt efter om komponenten er tilsat eller ej.

1. Fit modellen og udfør modelkontrol. Du skal skitsere og kommentere de figurer du laver.
2. Angiv estimater for samtlige parametre i modellen, og giv en (kortfattet) fortolkning af de enkelte parametre.
3. Bestem et 95% prædiktionsinterval for styrken af en limning med lim hvor komponent A og C, men ikke B, er tilsat. Hvad er fortolkningen af prædiktionsintervallet?
4. Undersøg om komponent A og komponent B har samme effekt på styrken af limningen.
5. Firmaet har en formodning om at der er en synergieffekt ved at benytte komponent A og B samtidig, dvs. at der er en ekstra effekt i forhold til modellen ovenfor hvis både A og B tilsættes. Undersøg om dette er tilfældet.

   Vink: Opstil og undersøg en ny model for middelværdierne.