Københavns Universitet
Det Naturvidenskabelige Fakultet

Eksamen i Statistik 1

30. juni 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 tre opgaver med i alt 13 delspørgsmål. Alle delspørgsmål vægtes ens. Data til opgave 3 ligger i filen `rottevaekst.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 tal med alle $t_i > 0$, og lad $X_1, \ldots, X_n$ være uafhængige stokastiske variable hvor X_i er Poissonfordelt med middelværdi βt_i. Her er $\beta > 0$ en ukendt parameter der skal estimeres fra data.

1. Gør rede for at log-likelihoodfunktionen for en observation $x = (x_1, \ldots, x_n) \in \mathbb{N}_0^n$ er givet ved
$$\ell_x(\beta) = -\log L_x(\beta) = -S_x \log \beta + S_t \beta$$
hvor $S_x = \sum_{i=1}^n X_i$ og $S_t = \sum_{i=1}^n t_i$.
Bestem derefter scorefunktionen, informationsfunktionen og Fisherinformationen.

Definer to estimatorer, $\tilde{\beta}$ og $\check{\beta}$, på følgende måde:

$$\tilde{\beta} = \frac{1}{n} \sum_{i=1}^n \frac{X_i}{t_i} \quad \text{og} \quad \check{\beta} = \frac{\sum_{i=1}^n X_i}{\sum_{i=1}^n t_i}.$$

Bemærk at begge estimatorer er nul hvis alle $X_i = 0$ og således strengt taget ikke altid har værdier i parametermængden. Dette skal du ikke diskutere i spørgsmål 2.

2. Følgende to spørgsmål skal besvares for både $\tilde{\beta}$ og $\check{\beta}$:
   - Er estimatoren en central estimator for β?
   - Hvis ja: Er estimatoren en variansminimal central estimator for β?
3. Bestem maksimum likelihood estimatet for β når $S_x \neq 0$. Overvej desuden hvad der sker når $S_x = 0$.
4. Betragt datasættet bestående af følgende værdier:

```
> t
 [1] 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
> x
 [1] 1 3 0 1 0 1 6 4 3 2
```

Bestem et 95% konfidensinterval baseret på „den falske Wald-teststørrelse“.

Betragt desuden hypotesen $H : \beta = 2$ og udfør likelihood ratio testet for hypotesen. Du kan benytte det sædvanlige asymptotiske resultat for fordelingen af $LR(X) = -2\log Q(X)$ uden bevis.

5. I dette spørgsmål skal du ved hjælp af simulation undersøge momenter for $\tilde{\beta}$ og $\check{\beta}$. Mere præcist skal du

   - Simulere data fra modellen for $n = 10$ og $n = 20$. Den sande værdi af β er hele tiden 3, og $t_i = i/n$. For $n = 10$ kan dette gøres ved hjælp af kommandoerne

     ```
     t <- (1:10) / 10
     x <- rpois(10, trueBeta*t)
     ```

   - Beregne estimaterne $\tilde{\beta}$ og $\check{\beta}$ for hvert simuleret datasæt
   - Gentage dette 5000 gange og beregne gennemsnit og empirisk spredning af estimaterne
   - Sammenligne med de teoretiske værdier

   Besvarelse: Besvarelsen af spørgsmålet består af følgende ting:

   - En udfyldt version af nedenstående skema:

			Teoretiske værdier		Simulation	
Estimator	n	β	middelværdi	spredning	gennemsnit	spredning
$\tilde{\beta}$	10	3	**	**	**	**
	20	3	**	**	**	**
$\check{\beta}$	10	3	**	**	**	**
	20	3	**	**	**	**

   - Kommentarer til dine resultater

Opgave 2

Lad $X = (X_1, X_2, X_3, X_4)$ være en stokastisk variabel med værdier i $\mathbb{R}^4$ med følgende egenskaber:

- (X_1, X_2) og (X_3, X_4) er uafhængige
- (X_1, X_2) og (X_3, X_4) er begge normalfordelt på $\mathbb{R}^2$
- Middelværdierne er alle nul: $\mathrm{E}X_1 = \mathrm{E}X_2 = \mathrm{E}X_2 = \mathrm{E}X_4 = 0$
- Variansmatricerne for (X_1, X_2) og (X_3, X_4) er ens og givet ved

$$\mathrm{V}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \mathrm{V}\begin{pmatrix} X_3 \\ X_4 \end{pmatrix} = \begin{pmatrix} 1 & 3 \\ 3 & 9 \end{pmatrix}.$$

1. Bestem den simultane fordeling af X, og afgør om der er tale om en regulær normalfordeling på $\mathbb{R}^4$.

2. Svar på følgende spørgsmål (svarene skal begrundes):

   - Har X_2 og X_4 samme fordeling?
   - Har $X_1 + X_2$ og $X_1 + X_4$ samme fordeling?
   - Findes der et underrum U af R^2 således at $\dim(U) = 1$ og

$$P\Big((X_1 + X_2\,, X_1 + X_4) \in U\Big) = 1?$$

Opgave 3

Data til denne opgave stammer fra et rotteforsøg med et væksthormon. Otteogtyve rotter blev inddelt i syv lige store grupper svarende til forskellige doser af væksthormonet. Forsøget varede en måned, og alle rotter blev vejet ved forsøgets start og afslutning. Vi skal interessere os for hvordan rotternes vækst, defineret som vægtøgningen i løbet af forsøget, afhænger af dosis.

Data er tilgængelige i filen `rottevaekst.txt` på den vedlagte USB-stick. Der er 28 datalinier og fire variable:

- `dosis`: Dosis af væksthormonet, varierer mellem 0 og 3.
- `dosisGrp`: Som variablen `dosis`, med bogstavet 'd' foran, således at variablen indlæses som en faktor
- `vaekst`: Vækst i gram (vægtøgning i løbet af forsøget)
- `logvaekst`: Den naturlige logaritme til variablen `vaekst`

Det anbefales at starte med at lave et (`dosis`, `vaekst`) scatterplot og et (`dosis`, `logvaekst`) scatterplot for at få en fornemmelse for data.

1. Fit to ensidede variansanalysemodeller hvor `dosisGrp` benyttes som forklarende variabel. Du skal bruge `vaekst` henholdsvis `logvaekst` som responsvariabel i de to modeller.

   Lav residualplots for de to modeller, og gør rede for at modellen med `logvaekst` som responsvariabel er at foretrække. Du skal skitsere og kommentere plottene.

I det følgende skal du hele tiden bruge `logvaekst` som responsvariabel. Stokastiske variable svarende til `logvaekst` kaldes X nedenfor, således at EX er forventede værdier af log-vækst.

2. Benyt den ensidede variansanalysemodel fra spørgsmål 1 til at bestemme et estimat og et 95% konfidensinterval for følgende forskelle:

   - Forskellen i EX ved dosis 3 og dosis 0
   - Forskellen i EX ved dosis 3 og dosis 2

3. Fit en lineær regressionsmodel hvor EX afhænger lineært af dosis, og benyt modellen til at bestemme et estimat og et 95% konfidensinterval for følgende størrelser:

   - Forskellen i EX ved dosis 3 og dosis 0
   - EX ved dosis 0

4. Fit en kvadratisk regressionsmodel hvor EX er et andengradspolynomium af dosis, og benyt modellen til at bestemme et estimat og et 95% konfidensinterval forskellen i EX ved dosis 3 og dosis 0.

   Vink: Opskriv den ønskede forskel som en funktion af parametrene i modellen.

Den eneste forskel mellem de tre fittede modeller med `logvaekst` som respons er middelværdiunderrummene. Disse betegnes L_{anova} for den ensidede variansanalysemodel, L_{linreg} for den lineære regressionsmodel og L_{kvadreg} for den kvadratiske regressionsmodel.

5. Hvilke af underrummene L_{anova}, L_{linreg} og L_{kvadreg} er indeholdt i hinanden? Udfør modelreduktion, dvs. udfør relevante hypotesetests med henblik på at vælge den mest fornuftige model til data.

6. Betragt en rotte der får dosis 2.25. Vil en vægtøgning på 18 gram være usædvanlig? Bemærk at de 18 gram er ikke på log-skala.