Københavns Universitet
Det Natur- og Biovidenskabelige Fakultet

Eksamen i Statistik 2

23. juni 2016

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

Eksamenssættet består af tre opgaver med i alt 18 delspørgsmål. De tre opgaver vægtes ens. Data til opgave 3 ligger i filen `bus.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 en del af besvarelsen.

Opgave 1

1. For fast r, hvor r er et naturligt tal, betragt fordelingen med tæthed

$$f_p(x) = \binom{x+r-1}{x} p^x (1-p)^r \quad \text{for } x \in \mathbb{N}_0$$

mht tællemålet på $\mathbb{N}_0$. Fordelingen afhænger af parameteren $p \in (0,1)$.

Du kan uden bevis benytte at f_p er en tæthed. Det kan ligeledes benyttes uden bevis at der for $|a| < 1$ gælder

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

Lad $X_1, \ldots, X_n$ være uafhængige og identisk fordelte stokastiske variable med tæthed f_p, med kendt $r \in \mathbb{N}$ og ukendt $p \in (0,1)$.

(a) Opskriv likelihoodfunktionen og loglikelihoodfunktionen.

(b) Find scorefunktionen og informationsfunktionen. Find fortegnet på den forventede information.

(c) Gør rede for at der er en entydig maksimaliseringsestimator $\hat{p}$ og skriv den op.

(d) Sæt nu $r = 1$. Undersøg om $\hat{p}$ er konsistent.

(e) Sæt nu $r = 1$. Gør rede for at $\hat{p}$ er asymptotisk normalfordelt, og angiv parametrene i den asymptotiske fordeling.

Opgave 2

2. Betragt de to faktorer:

$$\begin{aligned} F &: \{1,\ldots,N\} \longrightarrow \{\text{F1, F2, F3, F4, F5}\} \\ G &: \{1,\ldots,N\} \longrightarrow \{\text{G1, G2, G3}\} \end{aligned}$$

Faktoren $F \times G$ antages at være surjektiv.

Betragt varianskomponentmodellen $X \sim \mathbf{N}(A\beta, \sigma^2\Sigma)$, hvor $A\beta \in L \subset \mathbb{R}^N$, $\dim(L) = k$, $\sigma^2 > 0$ og $\Sigma = I + \lambda BB^T$. Her er I identitetsmatricen og $\lambda \geq 0$. Vi antager yderligere at $A = A_G$ er designmatricen for faktorunderrummet for faktor G og matricen B er effektmatricen hørende til effektparret $(F, 1)$.

(a) Er X_i'erne uafhængige? (At svare ja eller nej er nok)

(b) Hvad er k? (Her skal både angives i ord hvad det er og angives en numerisk værdi)

(c) Antag at $\dim(F \times G) = N$ og at datasættet er ordnet efter faktor F, således at først kommer alle observationer med label $F1$ i faktor F, dernæst alle observationer med label $F2$, osv. Opskriv kovariansmatricen.

(d) Opskriv likelihoodfunktionen.

(e) Er der en anden estimator af parametrene i modellen end maksimaliseringsestimatoren, man kunne foretrække? Argumenter for dit svar.

Resten af spørgsmålene drejer sig ikke om varianskomponentmodellen ovenfor.

(f) Betragt de surjektive faktorer B og T, der antages at være usammenlignelige. De er begge forskellige fra den konstante faktor 1. Betragt deres tilhørende underrum L_B og L_T. Angiv hvilke af følgende udsagn, der er henholdsvis korrekte, falske eller ikke kan afgøres uden at vide mere om faktorerne.

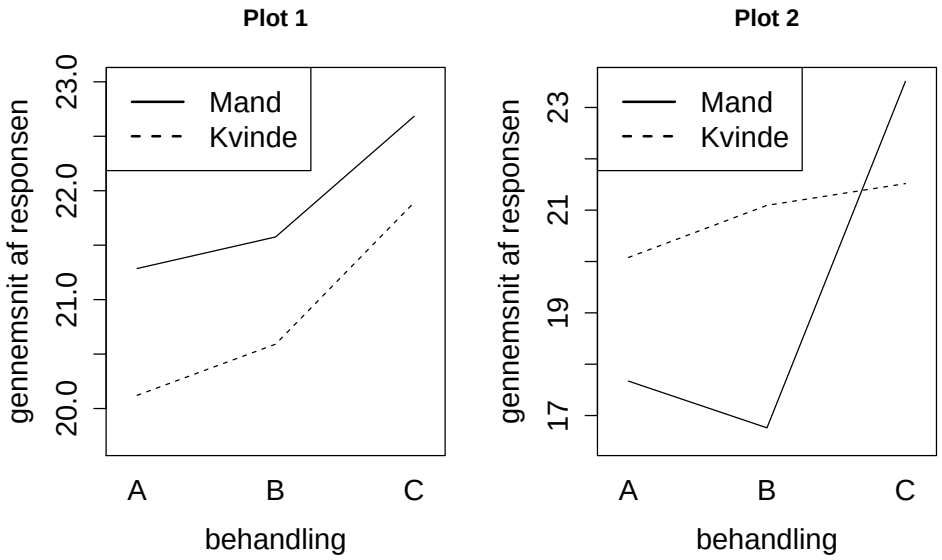
A. $L_B + L_T \subseteq L_{B\times T}$
B. $L_{B\times T} \subseteq L_B + L_T$
C. $L_{B\times T} \subseteq L_{B\wedge T}$
D. $L_{B\wedge T} \subseteq L_{B\times T}$
E. $L_B + L_T \subseteq L_{B\wedge T}$
F. $L_{B\wedge T} \subseteq L_B + L_T$
G. $L_B + L_T \subseteq L_1$

H. $L_1 \subseteq L_B + L_T$
I. $L_1 \subseteq L_{B \wedge T}$
J. $L_{B \wedge T} \subseteq L_1$

(g) Lad L_1 og L_2 være to underrum, begge forskellige fra $\{0\}$. Hvilke af følgende udsagn er korrekte?

A. Hvis $L_1 \underset{G}{\perp} L_2$ så er $L_1 \subset L_2$
B. Hvis $L_1 \subset L_2$ så er $L_1 \underset{G}{\perp} L_2$
C. Hvis $L_1 \subset L_2$ så er $L_1 \perp L_2$
D. Hvis $L_1 \perp L_2$ så er $L_1 \underset{G}{\perp} L_2$
E. Hvis $L_1 \underset{G}{\perp} L_2$ så er $L_1 \perp L_2$

(h) Betragt følgende interaktionsplots mellem de to faktorer `Behandling` og `Køn`, med henholdsvis 3 og 2 kategorier.



i. Vurder for hvert af ovenstående to interaktionsplots om de bedst beskrives med en vekselvirkningsmodel eller med en additiv model.

ii. Antag at responsen er lungekapacitet, og at man gerne vil have at den er stor. Hvilken behandling bør anbefales i hvert tilfælde?

iii. Antag at responsen er blodtryk, og at man gerne vil have at den er lille. Hvilken behandling bør anbefales i hvert tilfælde?

Opgave 3

3. Ved en undersøgelse af virkningen af forskellige dæktyper på benzinforbruget af offentlige busser blev følgende forsøg gennemført: 3 busser, A, B og C gennemkørte adskillige gange samme rundstrækning på ca. 10 km med 3 forskellige dæktyper K, L og M, og benzinforbruget i milliliter blev målt.

   Data er tilgængelige i filen `bus.txt` og består af variablene `bus`, `daek` og `benzin`, hvor den sidste angiver benzinforbruget.

   Vi antager i det følgende at de målte benzinforbrugstal kan ses som realisationer af uafhængige, normalfordelte stokastiske variable med samme varians σ^2 og med en middelværdi der potentielt afhænger af bussen og dæktypen. Vi indicerer observationerne ved mængden I, og betragter to faktorer:

$$\begin{aligned} \texttt{Bus} &: I \longrightarrow \{A, B, C\} \\ \texttt{Dæk} &: I \longrightarrow \{K, L, M\} \end{aligned}$$

   I spørgsmålene nedenfor bør angives relevante kvadrerede projektionslængder, dimensioner, F-test størrelser og fordelinger, både teoretisk og med numeriske værdier.

   (a) Gør rede for at de to faktorer er geometrisk ortogonale og opstil en passende statistisk model for data.

   (b) Undersøg om der er en signifikant vekselvirkning mellem de to faktorer.

   (c) Fortsæt med den additive model. Undersøg om der er en signifikant forskel på de tre bussers benzinforbrug. Test om dæktypen påvirker benzinforbruget.

   (d) Estimer parametrene i den additive model hvor begge faktorer indgår, og angiv deres simultane fordeling.

   (e) De to busser A og B er samme mærke bus, hvorimod bus C er af et andet mærke. Dermed kan det tænkes at busserne A og B virker ens. Opstil og test denne hypotese.