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.

Solution: Likelihoodfunktion:

$$\begin{aligned} L_X(p) = \prod_{i=1}^{n} f_p(X_i) &= \prod_{i=1}^{n} \left[\binom{X_i + r - 1}{X_i} p^{X_i} (1-p)^r \right] \\ &= p^{\sum_{i=1}^{n} X_i} (1-p)^{nr} \prod_{i=1}^{n} \binom{X_i + r - 1}{X_i} \end{aligned}$$

(Minus) loglikelihoodfunktion:

$$l_X(p) = -\log L_X(p) = -\sum_{i=1}^{n} \log \binom{X_i + r - 1}{X_i} - \sum_{i=1}^{n} X_i \log p - nr \log(1-p)$$

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

Solution: Scorefunktion:

$$\frac{d}{dp} l_X(p) = -\frac{1}{p} \sum_{i=1}^{n} X_i + \frac{nr}{1-p}$$

Informationsfunktion:

$$\frac{d^2}{dp^2} l_X(p) = \frac{1}{p^2} \sum_{i=1}^{n} X_i + \frac{nr}{(1-p)^2}$$

Da $X_i \geq 0$ for alle $i = 1, \ldots, n$ er informationsfunktionen åbenlyst positiv.

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

Solution: Likelihoodligningen findes ved at sætte scorefunktionen lig nul. Informationsfunktionen er skarpt positiv for alle $0 < p < 1$. En eventuel løsning til likelihoodligningen vil derfor være et globalt minimum for $l_X(p)$. Likelihoodligningen er

$$\frac{1}{p} \sum_{i=1}^{n} X_i = \frac{nr}{1-p}$$

hvis løsning giver maksimaliseringsestimatoren

$$\hat{p} = \frac{\frac{1}{n} \sum_{i=1}^{n} X_i}{r + \frac{1}{n} \sum_{i=1}^{n} X_i}$$

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

Solution: Lad $r = 1$. Da er

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

og $f_p(x) = p^x(1-p)$ for $x \in \mathbb{N}_0$ Vi skal bruge $E(X)$:

$$E(X) = \sum_{x=0}^{\infty} x p^x (1-p) = \frac{p}{1-p}$$

der ses at eksistere (sum af positive led), så LLN giver at

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \frac{p}{1-p}.$$

Vi har at

$$\hat{p} = f\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \quad \text{hvor} \quad f(x) = \frac{x}{1+x}.$$

Da $f(x)$ er kontinuert for alle $x > 0$ (og specielt i $f(p)$) kan (5.3) benyttes, og vi får at

$$\hat{p} \xrightarrow{P} f\left(\frac{p}{1-p}\right) = p.$$

Således er $\hat{p}$ konsistent.

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

Solution: Vi skal også bruge $\mathrm{Var}(X)$. Først ses at andetmomentet eksisterer:

$$E(|X|) = \sum_{x=0}^{\infty} x^2 p^x (1-p) = \frac{p(1+p)}{(1-p)^2} < \infty$$

Vi får

$$\mathrm{Var}(X) = \frac{p(1+p)}{(1-p)^2} - \frac{p^2}{(1-p)^2} = \frac{p}{(1-p)^2}$$

CLT giver da at

$$\frac{1}{n}\sum_{i=1}^{n} X_i \overset{as}{\sim} N\left(\frac{p}{1-p}, \frac{1}{n}\frac{p}{(1-p)^2}\right)$$

Vi benytter nu deltametoden da f er differentiabel for $p \in (0,1)$:

$$\hat{p} = f\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \overset{as}{\sim} N\left(f\left(\frac{p}{1-p}\right), \frac{1}{n}\left(f'\left(\frac{p}{1-p}\right)\right)^2 \frac{p}{(1-p)^2}\right)$$

således at

$$\hat{p} \overset{as}{\sim} N\left(p, \frac{1}{n}p(1-p)^2\right)$$

Opgave 2

2. Betragt de to faktorer:

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

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)

Solution: Nej hvis $\lambda > 0$. Ja hvis $\lambda = 0$.

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

Solution: $k = 3$ er dimensionen af middelværdiunderrummet.

(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.

Solution: $\mathrm{Var}(X) = \sigma^2\Sigma$ hvor

$$\sigma^2\Sigma = \sigma^2 \begin{pmatrix} \Sigma_1 & 0 & 0 & 0 & 0 \\ 0 & \Sigma_2 & 0 & 0 & 0 \\ 0 & 0 & \Sigma_3 & 0 & 0 \\ 0 & 0 & 0 & \Sigma_4 & 0 \\ 0 & 0 & 0 & 0 & \Sigma_5 \end{pmatrix}$$

og for $j = 1, \dots, 5$ er

$$\Sigma_j = \begin{pmatrix} 1+\lambda & \lambda & \lambda \\ \lambda & 1+\lambda & \lambda \\ \lambda & \lambda & 1+\lambda \end{pmatrix}$$

(d) Opskriv likelihoodfunktionen.

Solution: Likelihoodfunktion:

$$L_X(\beta, \sigma^2, \lambda) = \left(\frac{1}{2\pi\sigma^2}\right)^{N/2} \frac{1}{\sqrt{|\Sigma|}} \exp\{-(X - A\beta)^T \Sigma^{-1} (X - A\beta)/2\sigma^2\}$$

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

Solution: Ja, MLE underestimerer variansparametrene, især λ underestimeres. I stedet benyttes REML.

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. $L_{B\wedge T} \subseteq L_1$

Solution: A. Sand
B. Kan ikke afgøres. Det er næsten altid falsk, men her er et eksempel hvor det er sandt: 2 kategorier for hver faktor, 3 observationer, med antalstabel:

1	1
1	0

Her er $B \wedge T = 1$ (en sammenhængskomponent i designgrafen) så $\dim(L_B + L_T) = \dim(L_B) + \dim(L_T) - \dim(L_{B\wedge T}) = 2 + 2 - 1 = 3$ og fra antalstabellen ses at $\dim(L_{B\times T}) = 3$. Da de har samme dimension og $L_B + L_T \subseteq L_{B\times T}$ må $L_B + L_T = L_{B\times T}$
C. Falsk

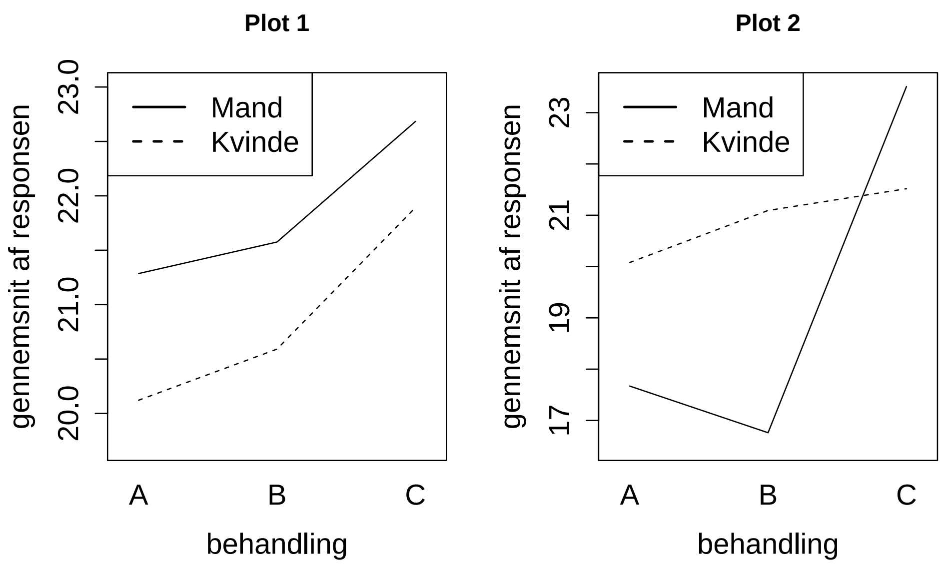
D. Sand
E. Falsk (da de er usammenlignelige, og derfor ikke kan være ens)
F. Sand
G. Falsk
H. Sand
I. Sand
J. Det kan man ikke afgøre, det afhænger af $\dim(L_{B \wedge T})$

(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 $L_1 \perp L_2$

Solution: A. Falsk
B. Sand
C. Falsk
D. Sand
E. Falsk

(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.

> **Solution:** Plot 1: Additiv. Plot 2: Vekselvirkning.

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

> **Solution:** Plot 1: Behandling C. Plot 2: Behandling C.

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

> **Solution:** Plot 1: Behandling A. Plot 2: Behandling A for kvinder, behandling B for mænd.

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`, `dæk` 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.

Solution: Vi kalder faktorerne for B (`Bus`) og D (`Daek`). Faktorerne er geometrisk ortogonale hvis de opfylder balanceligningen Sætning 14.8 - alternativt kan lemma 13.11 benyttes da designet for de to faktorer er sammenhængende.
Antalstabel:

	K	L	M	
A	2	2	2	6
B	4	4	4	12
C	6	6	6	18
	12	12	12	36

Vi tjekker ligningerne i den første søjle, de øvrige er det samme:

$$2 = \frac{6 \cdot 12}{36}; \quad 4 = \frac{12 \cdot 12}{36}; \quad 6 = \frac{18 \cdot 12}{36}$$

Faktorerne er således geometrisk ortogonale.
Statistisk model for data:
X er regulært normalfordelt på $\mathbb{R}^I$ med middelværdi $\xi \in L_{B \times D}$ og varians $\sigma^2 I$,

hvor I er 36×36 identitetsmatricen.
Alternativt:

$$X_i \sim \mathcal{N}(\xi_i, \sigma^2)$$

uafhængige, hvor middelværdivektoren $\xi = (\xi_i)_{i \in I} \in L_{B \times D}$.

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

Solution: Hypotese: $\xi \in L_{B+D}$. Relevante størrelser (udregnet med formel (12.10)):

Faktor F	$\|\|P_F X\|\|^2$	dim L_F
I	329532176	36
$B \times D$	329139343	9
B	326914949	3
D	328516570	3
1	326308096	1

Dette giver yderligere følgende størrelser (formel (13.5), designet er ortogonalt og sammenhængende):

$$||P_{B+D}X||^2 = ||P_B X||^2 + ||P_D X||^2 - ||P_1 X||^2 = 329123424$$

$$\dim P_{B+D} = \dim P_B + \dim P_D - \dim P_1 = 5$$

F-test for vekselvirkning mellem de to faktorer:

$$\begin{aligned} F &= \frac{(||P_{B\times D}X||^2 - ||P_{B+D}X||^2)/(\dim L_{B\times D} - \dim L_{B+D})}{(||X||^2 - ||P_{B\times D}X||^2)/(\dim L_I - \dim L_{B\times D}))} \\ &= \frac{(329139343 - 329123424)/(9-5)}{(329532176 - 329139343)/(36-9)} = 0.2735 \end{aligned}$$

der skal vurderes i en $F(4, 27)$-fordeling. Vi får $p = 0.8925$ og accepterer derfor nulhypotesen om ingen 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.

Solution: For at undersøge om der er signifikant forskel på bussernes benzinforbrug, opstilles hypotesen om ingen forskel: $H_1 : \xi \in L_D$. F-teststørrelsen bliver:

$$\begin{aligned} F &= \frac{(||P_{B+D}X||^2 - ||P_D X||^2)/(\dim L_{B+D} - \dim L_D))}{(||X||^2 - ||P_{B+D}X||^2)/(\dim L_I - \dim L_{B+D})} \\ &= \frac{(329123424 - 328516570)/(5-3)}{(329532176 - 329123424)/(36-5)} = 23.01 \end{aligned}$$

der skal vurderes i en $F(2, 31)$-fordeling. Vi får $p < 0.00001$ og afviser derfor nulhypotesen om ingen virkning af behandling.

For at undersøge om dæktypen påvirker benzinforbruget, opstilles nulhypotesen: $H_2 : \xi \in L_B$. F-teststørrelsen bliver:

$$\begin{aligned} F &= \frac{(||P_{B+D}X||^2 - ||P_B X||^2)/(\dim L_{B+D} - \dim L_B))}{(||X||^2 - ||P_{B+D}X||^2)/(\dim L_I - \dim L_{B+D})} \\ &= \frac{(329123424 - 326914949)/(5-3)}{(329532176 - 329123424)/(36-5)} = 83.75 \end{aligned}$$

der skal vurderes i en $F(2, 31)$-fordeling. Vi får $p < 0.00001$ og afviser derfor nulhypotesen om ingen forskel mellem dæktyper.

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

Solution: Vi benytter Korollar 10.21 og formel (10.27) til σ^2. Her er A designmatricen, der er 36×5. Vi får (parametrisering med intercept)

$$(A^T A)^{-1} = \begin{bmatrix} 36 & 12 & 18 & 12 & 12 \\ 12 & 12 & 0 & 4 & 4 \\ 18 & 0 & 18 & 6 & 6 \\ 12 & 4 & 6 & 12 & 0 \\ 12 & 4 & 6 & 0 & 12 \end{bmatrix}^{-1} = \begin{bmatrix} 0.22 & -0.17 & -0.17 & -0.08 & -0.08 \\ -0.17 & 0.25 & 0.17 & 0.00 & 0.00 \\ -0.17 & 0.17 & 0.22 & 0.00 & 0.00 \\ -0.08 & 0.00 & 0.00 & 0.17 & 0.08 \\ -0.08 & 0.00 & 0.00 & 0.08 & 0.17 \end{bmatrix}$$

og $\hat{\beta} = (A^T A)^{-1} A^T X$, hvilket giver

$$\hat{\beta} = (2985.6, -181.8, 108.50, -255.0, 349.3)^T \text{ hvor } \hat{\beta} \sim \mathcal{N}\left(\beta, \sigma^2 (A^T A)^{-1}\right)$$

Her er β_1 intercept-estimatet, der angiver middelværdien af benzinforbruget for bus A med dæktype K, β_2 angiver det yderligere bidrag hvis det er bus B, β_3 angiver det yderligere bidrag hvis det er bus C, β_4 angiver det yderligere bidrag hvis det er dæktype L, og β_5 angiver det yderligere bidrag hvis det er dæktype M.

Hvis den studerende vælger parametrisering uden intercept fås

$$(A^TA)^{-1} = \begin{bmatrix} 6 & 0 & 0 & 2 & 2 \\ 0 & 12 & 0 & 4 & 4 \\ 0 & 0 & 18 & 6 & 6 \\ 2 & 4 & 6 & 12 & 0 \\ 2 & 4 & 6 & 0 & 12 \end{bmatrix}^{-1} = \begin{bmatrix} 0.22 & 0.06 & 0.06 & -0.08 & -0.08 \\ 0.06 & 0.14 & 0.06 & -0.08 & -0.08 \\ 0.06 & 0.06 & 0.11 & -0.08 & -0.08 \\ -0.08 & -0.08 & -0.08 & 0.17 & 0.08 \\ -0.08 & -0.08 & -0.08 & 0.08 & 0.17 \end{bmatrix}$$

og $\hat{\beta} = (A^TA)^{-1}A^TX$, hvilket giver

$$\hat{\beta} = (2985.6, 2803.8, 3094.1, -255.0, 349.3)^T \text{ hvor } \hat{\beta} \sim \mathcal{N}\left(\beta, \sigma^2(A^TA)^{-1}\right)$$

I tabelform fås følgende estimater for middelværdierne:

	Bus = A	Bus = B	Bus =C
Dæk = K	2985.6	2803.8	3094.1
Dæk = L	2730.6	2548.8	2839.1
Dæk = M	3334.8	3153.1	3443.3

Benzinforbruget estimeres derfor til at være mindst for bus B med dæktype L. Variansestimatet er

$$\tilde{\sigma^2} = \frac{||X||^2 - ||P_{B+D}X||^2}{(\dim L_I - \dim L_{B+D})} = 13185.56; \quad \tilde{\sigma^2} \sim \frac{\sigma^2}{31}\chi^2_{31}$$

Derudover er $\hat{\beta}$ og $\tilde{\sigma^2}$ uafhængige. Det er OK at angive estimaterne for ξ i stedet for β, men oversættelsen til noget fortolkeligt må da gerne følge med.

(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.

Solution: Vi definerer en ny faktor med to labels:

$$\text{B2 } (\texttt{Bus2}) \quad : \quad I \longrightarrow \{AB, C\}$$

Vi opstiller hypotesen $H_3 : \xi \in L_{B2+D}$. F-teststørrelsen bliver:

$$\begin{aligned} F &= \frac{(||P_{B+D}X||^2 - ||P_{B2+D}X||^2)/(\dim L_{B+D} - \dim L_{B2+D}))}{(||X||^2 - ||P_{B+D}X||^2)/(\dim L_I - \dim L_{B+D})} \\ &= \frac{(329123424 - 328991292)/(5-4)}{(329532176 - 329123424)/(36-5)} = 10.021 \end{aligned}$$

der skal vurderes i en $F(1, 31)$-fordeling. Vi får $p = 0.00346$ og afviser derfor nulhypotesen om at de to busser A og B har samme benzinforbrug.