Københavns Universitet
Det Naturvidenskabelige Fakultet

Eksamen i Statistik 2

21. juni 2018

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 fire opgaver og dækker både pensum fra 2017 og 2018. Hvis man ønsker at blive bedømt i forhold til pensum fra 2018 afleverer man besvarelser af opgave 1, 3, og 4. Hvis man ønsker at blive bedømt i forhold til gammelt pensum (2017) afleverer man besvarelser af opgaverne 2, 3, og 4.

Hvis man ønsker at blive bedømt efter gammelt pensum skal dette angives tydeligt på første side i besvarelsen.

Hvis intet er angivet, vil besvarelsen blive bedømt i forhold til pensum fra 2018 og kun besvarelser af opgave 1, 3, og 4 vil indgå i bedømmelsen. Angives, at besvarelsen ønskes bedømt i forhold til gammelt pensum, indgår kun besvarelser af opgaverne 2, 3, og 4 i bedømmelsen.

I begge tilfælde vil der være i alt 17 delspørgsmål, som skal bedømmes; alle delspørgsmål vægtes ens i bedømmelsen.

Data til Opgave 3 og 4 ligger på en USB-nøgle i filerne `moral.txt` og `koed.txt`. Nøglen skal afleveres tilbage når eksamen slutter, men udelukkende for at den kan genbruges. Den skal altså ikke indgå som del af besvarelsen.

Opgave 1

Bemærk: besvarelse af denne opgave bedømmes kun i forhold til 2018 (nyt) pensum.

Lad $X = (X_1,\, X_2,\, X_3,\, X_4)^T \in \mathbb{R}^4$ være normalfordelt $\mathcal{N}(\xi, \Sigma)$, hvor

$$\xi = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$$\Sigma = \frac{1}{2}\begin{pmatrix} 1 & -\frac{1}{2} & 0 & 0 \\ -\frac{1}{2} & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}$$

1. Er X regulært eller singulært normalfordelt? Begrund dit svar.
2. Lad $X = (X_1,\, X_2,\, X_3,\, X_4)^T$ være som ovenfor. Angiv om $Y = (X_1, X_2)^T$, hhv. $Z = (X_3, X_4)^T$ er regulært eller singulært normalt fordelte.
3. Angiv fordelingen af $(Y+Z)^T \Sigma_{Y+Z}^{-1}(Y+Z)$, hvor Σ_{Y+Z} er variansmatricen for $Y+Z$.

Bemærk at de næste delopgaver er uden sammenhæng med de foregående.

Lad nu $W \sim \mathcal{N}(\xi, \sigma^2 I_4)$ være regulært normal fordelt.

4. Opskriv design matricen A for følgende specifikation af ξ

$$\xi = A\beta = \begin{pmatrix} \beta_1 + \beta_3 \\ \beta_1 \\ \beta_2 + \beta_3 \\ \beta_2 \end{pmatrix}$$

5. Betragt en observation $w = (2,\, 5,\, 3,\, 1)^T$ fra W defineret som i spørgsmål 4).

   Find maksimum likelihood estimatorerne for β og σ^2 med designmatricen fra spørgsmål 4) og angiv standard errors for $\hat{\beta}_1, \hat{\beta}_2$ og $\hat{\beta}_3$.

   Du må her gerne benytte maksimaliserings-estimatoren for $\hat{\sigma}^2$, fremfor den centrale estimator til at beregne standard errors for $\hat{\beta}$.

Opgave 2

Bemærk: besvarelse af denne opgave bedømmes kun i forhold til 2017 (gammelt) pensum.

Betragt fordelingen med tæthed

$$f_\alpha(x) = \left(\frac{1}{\sqrt{2\pi}}\right) x^{-3/2} \exp\left\{-\frac{(x-\alpha)^2}{2\alpha^2 x}\right\} \qquad \text{for } x > 0$$

med hensyn til Lebesgue-målet. Fordelingen afhænger af parameteren $\alpha > 0$.

Du kan uden bevis benytte at f_α er en tæthed.

1. Reparametriser f_α-fordelingen ved $\theta = -\frac{1}{2\alpha^2}$ så det fremgår at det er en eksponentiel familie med kanonisk stikprøvefunktion $t(x) = x$.
2. Identificer grundmålet. Identificer normeringskonstanten $c(\theta)$.
3. Lad X have tæthed f_α. Argumenter for at X har momenter af enhver orden. Find middelværdi og varians af X, både som funktion af θ og som funktion af α.
4. Opskriv likelihoodligningen for θ. Find maksimaliseringsestimatoren for θ. Find maksimaliseringsestimatoren for α.
5. Gør rede for at $\hat{\theta}$ er asymptotisk normalfordelt, og angiv parametrene i den asymptotiske fordeling, parametriseret ved θ.

Opgave 3

I et studie[1] undersøgte man om der var forskel på mænd og kvinders moral i den amerikanske kystvagt. Specifikt noterede man en score for hver person udfra den såkaldte *Rest's Defining Issues Test (DIT)*, hvor en højere score indikerer en højere moral.

Studiet involverede 225 personer fordelt på 4 grupper efter køn og rang. Tabel 1 viser antallet af personer i hver af de fire grupper. Data til besvarelse af opgaven findes på USB nøglen med filnavnet `moral.txt`.

		Rang	
		Officer	Menig
Køn	Mand	60	120
	Kvinde	15	30

Tabel 1: Antalstabel for studiet omkring mænd og kvinders moral i US Coast Guard.

Betragt faktorerne $K = \{\text{mand}, \text{kvinde}\}$ og $R = \{\text{officer}, \text{menig}\}$ med hver to niveauer.

1. Gør rede for at de to faktorer K og R er geometrisk ortogonale og angiv hvorvidt designet

$$\mathbb{G} = \{K, R, K \times R\}$$

er ortogonalt og minimums-stabilt.

2. Find $\dim(V_G)$ og $||Q_G X||^2$ for hver faktor G i faktorstrukturdiagrammet hvor V_G er den ortogonale dekomposition

$$V_G = L_G \ominus \sum_{G' \in \mathbb{G}, G' < G} L_{G'}, \quad \text{for hvert } G \in \mathbb{G}$$

med tilhørende projektion Q_G, og angiv $\dim L_K + L_R$.

[1] Kilde: R.D. White, Jr. (1999). "*Are Women More Ethical? Recent Findings on the Effects of Gender Under Moral Development*,"Journal of Public Administration Research and Theory, Vol. 9, #3, pp.459-471

Værdierne i diagrammet (1) kan benyttes.

$$
\begin{tikzcd}
 & & K^{252453,\,2} \arrow[dr] & \\
I^{288373,\,225} \arrow[r] & K \times R^{255038,\,4} \arrow[ur] \arrow[dr] & & 1^{251710,\,1} \\
 & & R^{254250,\,2} \arrow[ur] &
\end{tikzcd}
\tag{1}
$$

3. Idet vi antager en lineær normal model

$$X \sim \mathcal{N}(\xi, \sigma^2 I_{225})$$

hvor I_{225} betegner den 225-dimensionale identitetsmatrix, udfør F-testet der sammenligner modellerne $\xi \in L_{K \times R}$ og $\xi \in L_K + L_R$, udfra projektionerne Q_G fra spørgsmål 2).

4. Angiv de estimerede middelværdier for de fire grupper, estimeret udfra modellen

$$X \sim \mathcal{N}(\xi, \sigma^2 I_{225}),$$
$$\text{hvor } \xi \in L_{K \times R}$$

og undersøg om residualerne kan antages at følge en normal-fordeling med konstant varians.

5. Test om det kan antages at der er vekselvirkning mellem faktorerne K og R med et signifikans niveau på 5% og forklar i ord hvad resultatet fortæller.

6. Antag nu modellen $\xi \in L_K + L_R$, dvs. modellen under H_0 i forrige spørgsmål, uanset din konklusion for H_0. Undersøg hvorvidt denne model kan reduceres yderligere.

7. Angiv konfidensintervaller for parametrene i modellen hvor $\xi \in L_K + L_R$ og beskriv hvad den fortæller om moralen for mænd og kvinder, samt menige og officerer i den amerikanske kystvagt.

Opgave 4

I et eksperiment med grisekød har man undersøgt hvordan kødets farve (intensitet af rød) falmer over tid som følge af to forskellige opbevaringer: B={lyst, mørkt}. Da farven på kød indikerer hvor attraktivt kødet vurderes af forbrugerne, er man interesseret i hvordan de to behandlinger påvirker kødets farve, og specielt hvorvidt der er forskel på de to behandlinger.

I eksperimentet målte man på 10 slagtede grise. Fra hver gris tog man 6 stykker kød, hvoraf 3 blev lagret lyst i hhv. 1,4 og 6 dage og 3 blev lagret mørkt i hhv. 1,4 og 6 dage. Dette giver i alt $10 \times 3 \times 2 = 60$ observationer. Tabel 2 viser designet for hver gris.

Opbevaring	1 dag	4 dage	6 dage
Lyst	Stykke 1	Stykke 2	Stykke 3
Mørkt	Stykke 4	Stykke 5	Stykke 6

Tabel 2: Fordeling af kødstykker fra hver gris.

Figur 1 viser et plot af observationerne fordelt på de to opbevaringsmetoder.

Som det fremgår af Tabel 3 er der ialt 3 faktorer i spil.

Faktor	Betydning	Niveauer
G	Gris	$\{1, \ldots, 10\}$
T	Tid	$\{1, 4, 6\}$
B	Opbevaring	{lyst, mørkt}

Tabel 3: Faktorer i eksperimentet med grisekød.

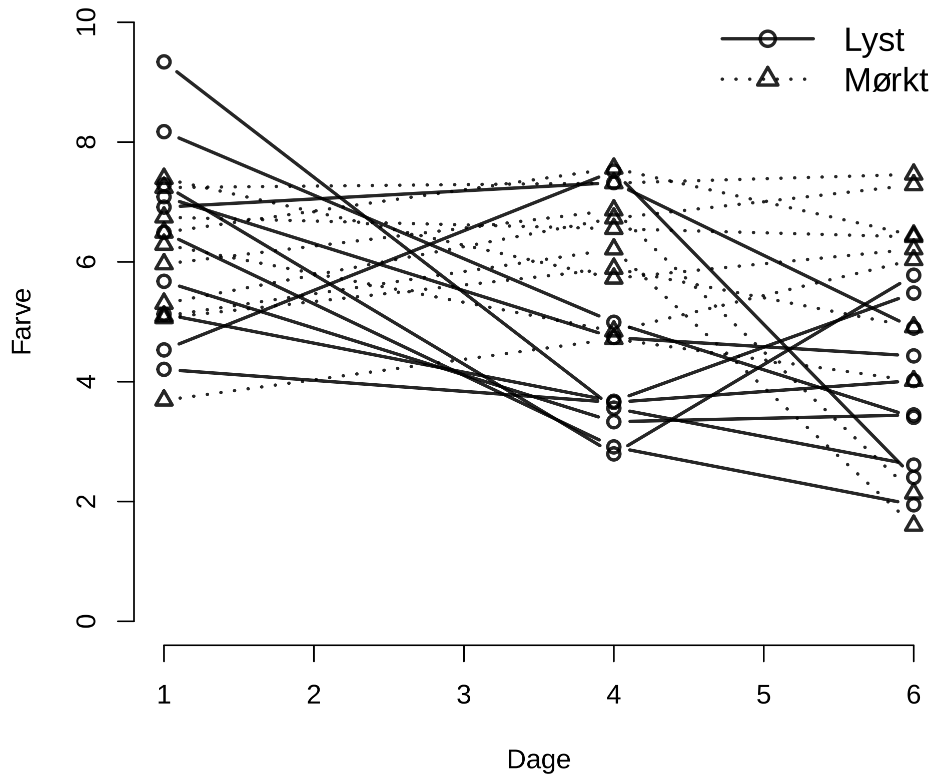
Betragt G som tilfældig effekt og T, B som faste effekter. Betragt derudover T (tid) som faktor og ej numerisk.

Data til besvarelse af opgaven findes på USB nøglen med filnavnet `koed.txt`.

Lad $i = 1, 2,\ j = 1, 2, 3,\ k = 1, \ldots, 10$ angive i'te behandling (opbevaring) til j'te tidspunkt for den k'te gris.

1. Opskriv en varianskomponentmodel for rødheden af kødet, med $T \times B$ (vekselvirkning mellem T og B) som fast effekt og G som tilfældigt intercept. Angiv variansmatricen for en vilkårlig gris udtrykt ved hjælp af de teoretiske parametre i modellen.

2. Angiv estimaterne for variansparametrene i modellen fra spørgsmål 1). Husk at variablerne `gris` og `tid` skal betragtes som faktorer og ej numeriske!



Figur 1: Effekt af 2 forskellige opbevaringer af grisekød til 3 forskellige tidspunkter.

3. Angiv de estimerede middelværdier for rødheden af kød for de tre $B \times T$-niveauer:

$$\{\text{lyst, 1 dag }\}$$
$$\{\text{mørkt, 1 dag }\}$$
$$\{\text{mørkt, 6 dage }\}$$

udfra estimaterne fra modellen fra spørgsmål 1). Bemærk at der spørges til estimerede middelværdier for niveauerne i produktfaktoren, snarere end parameterestimater. Sidstnævnte afhænger af den parametrisering du har valgt, men det gør middelværdierne ikke!

4. Opskriv konfidensintervaller for middelværdierne tilhørende de faste effekter i modellen med vekselvirkning, dvs. hvor $\xi \in L_{T\times B}$. Kommenter på effekten af de to behandlinger.

5. Angiv et forslag til at teste hvorvidt tid har nogen effekt når kødet er opbevaret mørkt. Du behøver ikke at udføre testet.