Københavns Universitet
Institut for Matematiske Fag

Eksamen i Matematisk Statistik
18. juni 2020

Eksamen er en 8 timers individuel skriftlig opgave. Alle hjælpemidler er tilladte. Besvarelsen skal udarbejdes alene og afleveres i Digital Eksamen som en samlet pdf-fil inkl. R-kode.

Opgavesættet består af 4 opgaver med i alt 14 delopgaver. Ved bedømmelsen vægtes alle delopgaver ens. Filen `MatStat2020Juni_opg3.txt` *indeholder data til opgave 3, og den er tilgængelig via Digital Eksamen såvel som på Absalon.*

Opgave 1

Lad $X_1,\ldots,X_n$ og $Y_1,\ldots,Y_n$ være indbyrdes uafhængige og eksponentialfordelte med $\mathbf{E}(X_i)=\beta$ og $\mathbf{E}(Y_i)=\gamma$ for $i=1,\ldots,n$, hvor $(\beta,\gamma)\in(0,\infty)^2$ er ukendte, og betragt hypotesen $H_0:\gamma=\beta^2$.

1. Gør rede for, at H_0 specificerer en en-dimensional krum eksponentiel familie af orden 2.

2. Opskriv log-likelihood funktionen, scorefunktionen og informationsfunktionen for β under antagelse af H_0.

3. Vis at maksimaliseringsestimatoren (MLE) $\hat{\beta}_n$ for β under hypotesen er entydigt bestemt som
$$\hat{\beta}_n=\frac{\bar{X}_n+\sqrt{\bar{X}_n^2+24\bar{Y}_n}}{6}$$
hvor $\bar{X}_n=(X_1+\cdots+X_n)/n$ og $\bar{Y}_n=(Y_1+\cdots+Y_n)/n$.

4. Angiv den asymptotiske fordeling af MLE for β.

5. I en stikprøve med $n=10$ observationer som ovenfor, opnåede man værdierne $\bar{X}_{10}=2.15$ og $\bar{Y}_{10}=5.349$.
   (a) Kan hypotesen $H_0:\gamma=\beta^2$ opretholdes på baggrund af disse observationer?

   (b) Kan hypotesen $H_1:\beta=1$ opretholdes under antagelse af H_0?

Opgave 2

Lad X og Y være uafhængige og identisk poissonfordelte stokastiske variable med middelværdi e^θ for $\theta \in \mathbb{R}$. Sæt $Z = X - Y$, og lad $t : \mathbb{Z} \to \mathbb{R}$ være $t(z) = z^2$ med momentfunktion

$$m(\theta) = E_\theta(t(Z)).$$

Det kan i opgaven uden bevis benyttes, at

$$E_\theta(Z^4) = 2e^\theta + 12e^{2\theta}.$$

Vi observerer nu $Z_1, \ldots, Z_n$ som n uafhængige og identisk fordelte stokastiske variable, hvor Z_i for $i = 1, \ldots, n$ har samme fordeling som Z.

1. Vis at momentfunktionen er
$$m(\theta) = 2e^\theta$$
og find den tilhørende momentestimator, $\tilde{\theta}_n$, for θ.
2. Vis at $\tilde{\theta}_n$ er konsistent og asymptotisk normalfordelt. Find den asymptotiske varians.
3. Vælg passende værdier af n og θ, simuler fordelingen af $\tilde{\theta}_n$ og sammenlign med dens asymptotiske fordeling.

Opgave 3

Kromosomer er lange DNA-sekvenser, som findes i levende organismers celler. Et gen er en delsekvens af et kromosom, som cellerne oversætter til proteiner. En typisk organisme har titusindvis af gener fordelt på et mindre antal kromosomer. Organismer fra forskellige arter har ofte et stort antal gener tilfælles, dog kan to i øvrigt ens gener fra to arter godt have funktionelt ubetydelige forskelle såsom forskellige længder.

Nogle gener er ekstremt lange, og cellerne bruger meget tid og energi på at oversætte de lange gener til proteiner. Sådanne ekstremt lange gener bør være udsat for et evolutionært tryk i retning af at gøre dem kortere, men de skal på den anden side stadig opretholde deres funktion. Man kan nu forestille sig, at det evolutionære tryk virker på samme måde for alle arter, eller at det virker forskelligt for forskellige arter.

Denne opgave går ud på at undersøge hypotesen om at det evolutionære tryk virker på samme måde ved at udtrykke den som en additiv hypotese i en lineær normal model. I opgaven betragtes data fra 10 lange gener og 38 arter (alle vertebrater), og datasættet indeholder følgende fire variable.

- `L`. Genets længde.
- `G`. Et identifikationsnummer for genet.

- `S`. Den latinske betegnelse for arten.
- `C`. Den overordnede klasse, som arten tilhører.

Opgaven består af fire delopgaver. Delopgaverne 3 og 4 kan besvares uafhængigt af delopgaverne 1 og 2.

1. Gør rede for, ved brug af data, at `C` er en grovere faktor end `S` og bestem dimensionen af faktorrummet $L_{\mathtt{C}}$. Tegn faktorstrukturdiagrammet for designet

$$\mathbb{G} = \{1, \mathtt{C}, \mathtt{S}, \mathtt{G}, \mathtt{C} \times \mathtt{G}, \mathtt{S} \times \mathtt{G}\}.$$

2. Gør rede for, at designet $\mathbb{G}$ er stabilt overfor minimumsdannelse. Det kan uden bevis benyttes at $\mathtt{S} \wedge \mathtt{C} \times \mathtt{G} = \mathtt{C}$. Afgør om designet er ortogonalt, og find endelig dimensionerne af $L_{\mathtt{S}} + L_{\mathtt{G}}$ og $L_{\mathtt{S}} + L_{\mathtt{C}\times\mathtt{G}}$.
3. Fit en lineær normal model for længden `L` såvel som for $\sqrt[3]{\mathtt{L}}$, hvor middelværdirummet er givet ved

$$\mathtt{S} + \mathtt{C} \times \mathtt{G}.$$

   Afgør hvilken af de to modeller, der bedst fitter data.

4. Gør rede for, at

$$L_{\mathtt{S}} + L_{\mathtt{G}} \subseteq L_{\mathtt{S}} + L_{\mathtt{C}\times\mathtt{G}},$$

   og test den additive hypotese

$$H_0 : \mathtt{S} + \mathtt{G}.$$

Opgave 4

Lad $X_1, \ldots, X_n$ are uafhængige og identisk normalfordelte variable med værdier i $\mathbb{R}^3$, middelværdi 0 og variansmatrix

$$V(X_i) = \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{pmatrix}$$

for $i = 1, \ldots, n$. Betragt endvidere hypotesen

$$H_0 : \Sigma = \begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{pmatrix}.$$

I det følgende benytter vi notationen

$$\mathrm{S} = \begin{pmatrix} \mathrm{S}_{11} & \mathrm{S}_{12} & \mathrm{S}_{13} \\ \mathrm{S}_{12} & \mathrm{S}_{22} & \mathrm{S}_{23} \\ \mathrm{S}_{13} & \mathrm{S}_{23} & \mathrm{S}_{33} \end{pmatrix} = \sum_{i=1}^{n} X_i X_i^T.$$

Det kan endvidere antages, at S er positiv definit.

1. Gør rede for at maksimaliseringsestimatoren for Σ er $\hat{\Sigma} = \frac{1}{n}\mathrm{S}$, og at maksimaliseringsestimatoren under hypotesen H_0 er

   $$\hat{\sigma}_1^2 = \frac{1}{n}\mathrm{S}_{11}, \quad \hat{\sigma}_2^2 = \frac{1}{n}\mathrm{S}_{22}, \quad \hat{\sigma}_3^2 = \frac{1}{n}\mathrm{S}_{33}.$$

   Vis nu at kvotienttesstørrelsen for hypotesen H_0 er

   $$Q = \left(\frac{\det(\mathrm{S})}{\mathrm{S}_{11}\mathrm{S}_{22}\mathrm{S}_{33}} \right)^{n/2}.$$

   Vink: Såvel model som hypotese kan ses som lineære hypoteser i en kendt eksponentiel familie.

2. Gør rede for at $-2\log Q$ er asymptotisk χ^2-fordelt og bestem antallet af frihedsgrader for den asymptotiske χ^2-fordeling. Beregn for $n = 30$ og

   $$\mathrm{S} = \begin{pmatrix} 40,51 & 4,92 & 22,84 \\ 4,92 & 34,55 & 16,39 \\ 22,84 & 16,39 & 47,46 \end{pmatrix}$$

   en p-værdi for test af H_0 og konkluder.