Københavns Universitet
Institut for Matematiske Fag

Eksamen i Matematisk Statistik
20. august 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 3 opgaver med i alt 14 delopgaver. Ved bedømmelsen vægtes alle delopgaver ens. Filerne `restaurant.txt` *og* `bodyfat.txt` *indeholder data til hhv. opgave 2 og opgave 3 og er tilgængelige via Digital Eksamen.*

Opgave 1

Vi siger at en positiv stokastisk variabel X er *logaritmisk normalfordelt* med parametre (ξ, σ^2), hvis X har samme fordeling som e^Y, hvor $Y \sim \mathcal{N}(\xi, \sigma^2)$. Man kan da vise, at X har tætheden

$$f_{(\xi,\sigma^2)}(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\log x - \xi)^2}{2\sigma^2}\right), \quad x > 0$$

med hensyn til standard Lebesgue mål på $\mathbb{R}$.

I det følgende kan det uden bevis benyttes at vi for alle $t > 0$ har

$$\int_{-\infty}^{\infty} e^{tz}\varphi(z)\,dz = \exp(t^2/2),$$

hvor φ betegner tætheden for standard normalfordelingen $\mathcal{N}(0,1)$.

Lad nu $X_1, \dots, X_n$ være en stikprøve af indbyrdes uafhængige og logaritmisk normalfordelte observationer med parametre (ξ, σ^2) hvor $\xi \in \mathbb{R}$ er ukendt mens σ^2 antages kendt.

1. Vis at middelværdien i den logaritmiske normalfordeling er
$$\mu = \mathbf{E}(X) = \exp(\xi + \sigma^2/2).$$
2. Opskriv log-likelihood funktionen, scorefunktionen og informationsfunktionen for μ baseret på en stikprøve $X_1 = x_1, \dots, X_n = x_n$ af observationer fra denne fordeling.
3. Bestem maximum likelihood estimatoren (MLE) $\hat{\mu}_n$ for middelværdien μ og find dens asymptotiske fordeling.

4. En stikprøve fra denne fordeling af størrelsen $n = 10$ gav følgende observationer

   `6.62  4.91  2.77 18.71  3.10  4.41 11.99  8.03  6.13  5.71`

   Idet det kan antages, at $\sigma^2 = 0.25$ ønskes det undersøgt, om observationerne understøtter hypotesen $H_0 : \mu = 6$.

5. Bestem momentestimatoren $\tilde{\mu}_n$ for μ baseret på momentfunktionen $t(x) = x$.

6. Bestem variansen for momentestimatoren $\tilde{\mu}_n$ og sammenlign med Cramér–Rao's nedre grænse.

7. Sammenlign estimatorerne $\hat{\mu}_n$ or $\tilde{\mu}_n$ ved simulation for passende parameterværdier i tilfældet med $n = 10$ og kommenter resultatet.

Opgave 2

En dansk restaurant holder åbent 10–15 på hverdage og 10–14 samt 17–22 i weekenden. Den serverer frokost alle ugens hverdage (mellem 10 og 15) samt brunch (mellem 10 og 14) og middag (mellem 17 og 22) på lørdage og søndage.

Ejeren af restauranten vil gerne undersøge effekten på omsætningen (X) af to forskellige online reklameplatforme. Hun er heldigvis uddannet statistiker og designer derfor et forsøg, hvor hun over 8 uger hver uge tilfældigt vælger at reklamere udelukkende på en af de to reklameplatforme den pågældende uge. Det resulterende datasæt i filen `restaurant.txt` indeholder følgende fire variable.

- X: Omsætningen i DKK.
- B: Serveringstypen på de tre niveauer: `fr, br, mi`.
- D: Ugedag på de syv niveauer: `m, ti, o, to, f, l, s`.
- T: Reklameplatform på de to niveauer: `fb og ta`.

1. Bestem dimensionen af $L_{\mathrm{T}\times\mathrm{B}}$ og find minimumsfaktoren $W = B \wedge D$. Gør rede for at designet

   $$\mathbb{G} = \{1, \mathrm{W}, \mathrm{T}, \mathrm{B}, \mathrm{D}, \mathrm{T} \times \mathrm{B}\}$$

   er stabilt overfor minimumsdannelse og tegn faktorstrukturdiagrammet.

2. Afgør hvorvidt $\mathbb{G}$ er et ortogonalt design og beregn dimensionerne af de to middelværdirum $L_{\mathrm{D}} + L_{\mathrm{T}\times\mathrm{B}}$ og $L_{\mathrm{D}} + L_{\mathrm{T}} + L_{\mathrm{B}}$.

3. Fit en lineær normal model for omsætningen X såvel som for $\log(X)$ med middelværdirum givet ved

   $$\mathrm{D} + \mathrm{T} \times \mathrm{B}.$$

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

4. Vis at
$$L_{\mathrm{D}}+L_{\mathrm{B}} \subseteq L_{\mathrm{D}}+L_{\mathrm{T}}+L_{\mathrm{B}} \subseteq L_{\mathrm{D}}+L_{\mathrm{T}\times\mathrm{B}},$$
og undersøg hvorvidt reklameplatformen giver anledning til en forskel i omsætningen.

5. Beregn et konfidensinterval for en parameter, der udtrykker forskellen i omsætningen mellem de to reklameplatforme. Fortolk resultatet.

Opgave 3

I denne opgave undersøges sammenhængen mellem menneskers fedtprocent og vægt. Datasættet i filen `bodyfat` indholder $n = 240$ observationer af samhørende værdier af vægt (i pund) og fedtprocent.

I opgaven kan det antages at observationerne er realiseringer af $n = 240$ uafhængige og identisk normalfordelte variable $X_1, \ldots, X_n$ med $X_i \sim \mathcal{N}(\mu, \Sigma)$, hvor $\mu \in \mathbb{R}^2$ og

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \lambda \\ \lambda & \sigma_2^2 \end{pmatrix}.$$

I første omgang betragtes hypotesen
$$H_0 : \lambda = 0.$$

1. Beregn for datasættet maksimaliseringsestimatorerne for μ og Σ under modellen såvel som under hypotesen H_0. Test hypotesen H_0.

Vi er herefter interesserede i hypotesen

$$H_1 : \lambda = \frac{1}{2}\sigma_1\sigma_2.$$

2 Vis at under H_1 er
$$\Sigma^{-1} = \frac{2}{3}\begin{pmatrix} \frac{2}{\sigma_1^2} & -\frac{1}{\sigma_1\sigma_2} \\ -\frac{1}{\sigma_1\sigma_2} & \frac{2}{\sigma_2^2} \end{pmatrix},$$
og gør rede for at hypotesen H_1 specificerer en krum eksponentiel familie af dimension 4 og orden 5.