

Sandsynlighedsteori og Statistik: Ugeseddel 10

Levi van Boekel

Uge 46, 2024

Opgave 1

1.1

Idet vi antager, at udfaldene er observationer, y_i , af stokastiske variable, Y_i , $i = 1, \dots, 21$, fra en Poisson fordeling, er udfaldsrummet for de stokastiske variable blot alle ikke negative tal, idet Poisson fordelingen er defineret for alle ikke negative tal. Dvs:

$$Y_i \in \mathbb{Y} = \mathbb{N} \quad (1)$$

Parameteren i Poisson fordeling kan tiltage alle værdier større end 0, da har vi parameterrummet givet som:

$$\lambda_i \in \Theta = \{\lambda \in \mathbb{R} : \lambda > 0\} \quad (2)$$

1.2

Hvis de stokastiske variable er identisk fordelt, kan vi blot skrive den generelle Poisson sandsynlighedsfunktion og evaluere den i de forskellige observationer. Altså:

$$f_{Y_i}(y_i | \lambda) = \frac{e^{-\lambda} \lambda^{y_i}}{y_i!} \quad y_i \in \mathbb{N} \quad (3)$$

Ud fra oplysningerne i opgaven er det svært at sige, om det er en rimelig antagelse, men det er klart, at diskussionen ændrer sig, hvis nogle virksomheder er fra service branchen og andre er fra IT-branchen, idet sandsynligheden for at udtage et patent må siges at være større for en virksomhed fra IT-branchen. I dette tilfælde vil man med fordel kunne opstille en betinget model.

1.3

Definitionen på, at to stokastiske variable er, at den simultane ssh. er lig produktet af de marginale, for alle $i \neq j$ skal det da gælde, at:

$$P(x_i, y_i) = P(x_i) \cdot P(y_i)$$

Hvis vi antager, at der for alle $i = 1 \dots 21$ gælder indbyrdes uafhængighed bliver den simultane sandsynlighedsfunktion blot produktet af de marginale:

$$f_{Y_1, \dots, Y_{21}}(y_1 \dots y_{21} \mid \lambda) = \prod_{i=1}^{21} \frac{e^{-\lambda} \lambda^{y_i}}{y_i!} \quad y_i \in \mathbb{N}, \lambda \in \Theta \quad (4)$$

Det forekommer rimeligt at antallet af patenter en virksomhed udtager er uafhængighed af hvor mange patenter andre virksomheder udtager (givet at virksomhederne er tilfældigt udvalgt).

1.4

Sample likelihood funktionen er givet ved:

$$L(\lambda \mid y_1, \dots, y_{21}) = \prod_{i=1}^{21} \frac{e^{-\lambda} \lambda^{y_i}}{y_i!} \quad y_i \in \mathbb{N}, \lambda \in \Theta$$

Værdien af likelihood funktionen kan tolkes som den simultane ssh. for at få realisationerne med den givne fordeling (Poisson) og den indsatte parameter værdi. Vi kan tage logaritmen og få *log-likelihood funktionen*, hvor produktet erstattes af summer. Det gælder, da logaritmefunktionen er en monotont voksende transformation og dermed ikke maksimumsværdien af en funktion (og det netop er den vi er interesseret i at finde).

$$\log L(\lambda \mid y_1 \dots y_{21}) = \sum_{i=1}^{21} [y_i \cdot \log(\lambda) - \lambda - \log(y_i!)] \quad (5)$$

Bemærk, at vi ved at finde maksima af ovenstående kan finde *MLE estimatet*. Hvis vi vil beskrive estimatet ved hjælp af statistiske metoder, skal vi finde opstille *MLE estimatoren*, der er en stokastisk variabel i sig selv og da kan beskrives ved de metoder vi kender fra stokastiske variable. Log likelihood funktionen med stokastiske variable er da:

$$\log L(\lambda \mid Y_1 \dots Y_{21}) = \sum_{i=1}^{21} [Y_i \cdot \log(\lambda) - \lambda - \log(Y_i!)] \quad (6)$$

1.5

Første ordens betingelsen er:

$$\begin{aligned}\frac{\delta \log L(\lambda \mid Y_1, \dots, Y_{21})}{\delta \lambda} &= 0 \\ \sum_{i=1}^{21} \left[\frac{Y_i}{\lambda} - 1 \right] &= 0 \\ \frac{\sum_{i=1}^{21} Y_i}{\lambda} &= n \\ \sum_{i=1}^{21} Y_i &= \lambda n \\ \lambda &= \frac{1}{n} \sum_{i=1}^{21} Y_i \equiv \hat{\lambda}(Y_1, \dots, Y_{21})\end{aligned}$$

Som blot er det aritmetriske gennemsnit.

1.6

Den anden ordens afledte er:

$$\begin{aligned}\left[\frac{\sum_{i=1}^{21} Y_i}{\lambda} - n \right]' &= 0 \\ -\frac{\sum_{i=1}^{21} Y_i}{\lambda^2} &< 0\end{aligned}$$

Sådan et Hessematrixen er negativ definit, og løsningen er et maksimum.

1.7

MLE estimatet er:

$$\hat{\lambda}(y_1, \dots, y_{21}) = \frac{1}{21} \cdot 63 = \frac{63}{21} = 3$$

1.8

Hvis ingen firmaer udtager patenter, bliver MLE estimatet $\lambda = 0$, men det er i modstrid med definitionsmængden for vores parameterrum, hvorfor Poissonfordelingen her ikke er defineret.

Opgave 2

2.1

Y_i er stokastiske variable, der måler levetiden af nye elpærer, mens y_i er realisationer af de stokastiske variable, der angiver en observation for levetiden af elpærerne. I tæthedsfunktionen angiver y således realisationer af stokastiske variable. Tæthedsfunktionen er skitseret nedenfor:

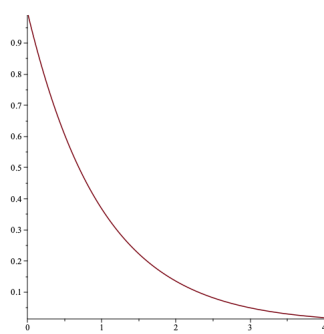


Figure 1: Tæthedsfunktionen med $\theta = 1$

2.2

Idet der ikke er noget fodtegn ved θ i tæthedsfunktionen kan vi se, at det antages, at dette ikke variere på tværs af de stokastiske variable. Antagelsen er nødvendig for at vi kan bruge MLE estimation, da vi ellers ikke ville kunne se udfaldende som realisationerne af ens fordelte stokastiske variable og da ikke kunne finde en parameterværdi, der maksimerer sandsynligheden for at få udfaldende, givet en fordeling.

2.3

Sample likelihood funktionen er:

$$L(\theta \mid y_1 \dots y_n) = \prod_{i=1}^{21} \theta \cdot e^{-\theta y_i} \quad y_i \in \mathbb{Y}, \theta \in \Theta \quad (7)$$

Hvor $\Theta = \{\theta \in \mathbb{R} : 0 < \theta < \infty\}$. Log likelihood sample likelihood funktionen er:

$$\log L(\theta \mid y_1 \dots y_n) = \sum_{i=1}^{21} [\log(\theta) - \theta y_i] \quad (8)$$

2.4

Det er vigtigt at antage uafhængighed, idet man ellers ikke kan opskrive den simultane sandsynlighed som produktet af de marginale. I.i.d antagelserne kan være problematiske, hvis levetiden

også afhænger af, hvordan husholdninger bruger deres pærer. Antagelsen om uafhængighed virker dog plausibel. Det er svært at sige, hvorvidt antagelsen om at den følger en eksponentielfordeling er rimelig.

2.5

MLE estimatoren findes til:

$$\begin{aligned}\sum_{i=1}^{21} \left[\frac{1}{\theta} - Y_i \right] &= 0 \\ \frac{n}{\theta} - \sum_{i=1}^{21} Y_i &= 0 \\ \hat{\theta}(Y_1, \dots, Y_n) &= \frac{n}{\sum_{i=1}^{21} Y_i}\end{aligned}$$

Den anden ordens afledte er:

$$\begin{aligned}\left[\frac{n}{\theta} - \sum_{i=1}^{21} Y_i \right]' &= 0 \\ -\frac{n}{\theta^2} &< 0\end{aligned}$$

Hvorfor Hessematricen er negativ definit og vi har et maksimum.

2.6

MLE estimatet findes ved at indsætte realisationer af de stokastiske variable:

$$\hat{\theta}(y_1, \dots, y_n) = \frac{121}{252} \approx 0.48$$

2.7

Forskellen er, at det ene er et estimat (en værdi) og det andet er estimator, dvs. en stokastisk variabel, der derfor kan beskrives ved de redskaber vi har for stokastiske variable (middelværdi, varians, osv). Hvis man læser, at $\hat{\theta} = 9.25$ menes der estimatet. Et udsagn som at $\hat{\theta}$ er god fordi den har lille varians, henviser derimod til estimatoren, da stokastiske variable — i modsætning til enkelte tal — har en varians.

Opgave 3

3.1

Givet, at vi betragter Y_i og Y_j og de stadigvæk er uafhængige for $i \neq j$, kan vi skrive sample likelihood funktionen som:

$$L(\theta)(Y_1 \dots, Y_{121}) = L(\theta_1)(Y_1 \dots, Y_{120}) \times L(\theta_2)(Y_{76} \dots, Y_{120}) = \prod_{i=1}^{75} \theta_1 e^{-\theta_1 y_i} \times \prod_{i=76}^{120} \theta_2 e^{-\theta_2 y_i}$$

$$\in y_i \in \mathbb{Y}, \theta \in \Theta$$

Hvor $\theta = (\theta_1, \theta_2)$ og $\Theta = \{\theta_1, \theta_2 \in \mathbb{R} : 0 < \theta_1, \theta_2 < \infty\}$. Log likelihood funktionen er:

$$\log L(\theta)(Y_1 \dots, Y_{121}) = \sum_{i=1}^{75} [\log(\theta_1) - y_i] + \sum_{i=76}^{120} [\log(\theta_2) - y_i]$$

3.2

MLE estimatoren er:

$$\hat{\theta}(Y_1, \dots, Y_{121}) = \begin{cases} \hat{\theta} = \frac{75}{\sum_{i=1}^{75} y_i} \\ \hat{\theta} = \frac{45}{\sum_{i=76}^{120} y_i} \end{cases}$$

3.3

MLE estimatet bliver da:

$$\hat{\theta}(y_1, \dots, y_{121}) = \begin{cases} \hat{\theta} = \frac{75}{160} \approx 0.46 \\ \hat{\theta} = \frac{45}{92} \approx 0.48 \end{cases}$$

MLE estimatet for de gamle pærer er højere end for de gamle. I vores model betyder det, at ssh. for, at en pære *dør* i hver periode er højere, hvorfor de gamle pærer er dårligere end de nye pærer.

3.4

Vi kan tolke den udvidede model som en betinget model, idet vi modificerer tæthedsfunktionerne alt afhængigt af, om det er nye eller gamle pærer. Hvis vi lod X angive, om pærerne var gamle med $x = 1$ (positiv) og $x = 0$ (negativ), har vi netop en betinget model. Hvor:

$$Y_i \mid X_i = x_i \sim \text{ekspontiel}(\theta_{x_i}) \quad (9)$$

Hvor $X_i \in \mathbb{I}$ og $\mathbb{I} = \{0, 1\}$.

Opgave 4

4.1

Idet et tal opløftet i 0 er 1 er det klart, at vi for $y \in \mathbb{Y}$ kan skrive gaffelfunktionen med hjælp af følgende indikatorfunktioner, der er 1, hvis $y = x$ og ellers er 0.

$$f_{Y_i}(y | \theta) = \left(\frac{2\theta}{3}\right)^{\mathbb{I}(y=0)} \cdot \left(\frac{\theta}{3}\right)^{\mathbb{I}(y=1)} \cdot \left(\frac{2(1-\theta)}{3}\right)^{\mathbb{I}(y=2)} \cdot \left(\frac{1-\theta}{3}\right)^{\mathbb{I}(y=3)} \quad (10)$$

4.2

For at finde sample likelihood funktionen sætter vi potenserne til det antal gange observationerne forekommer, dvs:

$$L(\theta | Y) = \left(\frac{2\theta}{3}\right)^2 \cdot \left(\frac{\theta}{3}\right)^3 \cdot \left(\frac{2(1-\theta)}{3}\right)^3 \cdot \left(\frac{1-\theta}{3}\right)^2 \quad (11)$$

Log likelihood bliver:

$$\begin{aligned} \log L(\theta | Y) &= 2 \cdot [\log(2\theta) - \log(3)] + 3 \cdot [\log(\theta) - \log(3)] + 3 \cdot [\log(2(1-\theta)) - \log(3)] \\ &\quad + 2 \cdot [\log(1-\theta) - \log(3)] \\ &= 5 \log(2) + 5 \log(\theta) - 10 \log(3) + 5 \log(1-\theta) \end{aligned}$$

4.3

Sample likelihood funktionen som funktion af θ er tegnet nedenfor:

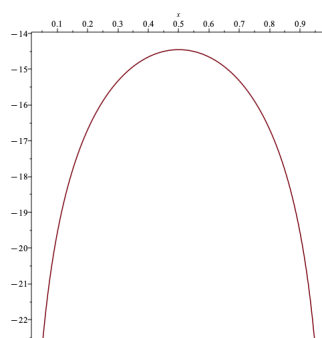


Figure 2: Sample likelihood funktion

4.4

MLE estimatet findes ved at finde score funktionen og sætte lig 0:

$$\begin{aligned} S &= \frac{5}{\theta} - \frac{5}{1-\theta} \\ \frac{5}{\theta} &= + \frac{5}{1-\theta} \\ 5 - 5\theta &= 5\theta \\ \hat{\theta} &= \frac{1}{2} \end{aligned}$$

4.5 & 4.6

Sandsynlighederne bliver:

$$\begin{array}{lll} P(Y_i = 0) = \frac{1}{3} & \text{observeret} & \frac{1}{5} \\ P(Y_i = 1) = \frac{1}{6} & \text{observeret} & \frac{3}{10} \\ P(Y_i = 2) = \frac{1}{3} & \text{observeret} & \frac{3}{10} \\ P(Y_i = 3) = \frac{1}{6} & \text{observeret} & \frac{1}{5} \end{array}$$

Modellen er altså nogleleunde i overensstemmelse med realisationerne.