

Sandsynlighedsteori og Statistik: Ugeseddel 11

Levi van Boekel

Uge 47, 2024

Opgave 1

1.1 & 1.2

Den empiriske fordeling af de første 5000 trækninger er beskrevet ved de første fire centrale momenter, samt percentiler nedenfor:

y				
	Percentiles	Smallest		
1%	2.601985	.875423		
5%	3.384351	1.35505		
10%	3.729487	1.505643	Obs	5,000
25%	4.336663	1.529272	Sum of wgt.	5,000
50%	4.985648		Mean	5.005519
		Largest	Std. dev.	1.009288
75%	5.683564	7.997656		
90%	6.324423	8.210836	Variance	1.018663
95%	6.703341	8.243766	Skewness	-.0094015
99%	7.314719	9.512827	Kurtosis	3.050507

Figure 1: Beskrivende statistik for de første 5000 trækninger

1.3

Vi ser, at gennemsnit går asymptotisk mod 5000, når $n \rightarrow \infty$. Det er et udtryk for LLN, der siger, at at det for $Y_1, \dots, Y_n$ forskellige ukorrelerede stokastiske variable med ens middelværdi (μ) og varians ($\sigma^2 < \infty$), for for ethvert $\delta > 0$, gælder, at:

$$P(|\frac{1}{n} \sum_{i=1}^n Y_i - \mu| \geq \delta) \rightarrow 0 \quad \text{når } n \rightarrow \infty \quad (1)$$

Opgave 2

2.1

Sandsynlighedsfunktionen, kan skrives som:

$$P(Y_i = y) = f_{Y_i}(y | p_1, p_2, p_3) = p_1^{\mathbb{I}(y=y_1)} \times p_2^{\mathbb{I}(y=y_2)} \times p_3^{\mathbb{I}(y=y_3)} \quad (2)$$

Fordi et tal opløftet i 0 er 1, og vi derfor, har:

$$f_{Y_i}(y \mid p_1, p_2, p_3) = \begin{cases} p_1 & \text{hvis } y = 1 \\ p_2 & \text{hvis } y = 2 \\ p_3 & \text{hvis } y = 3 \end{cases}$$

Som netop er det, der er givet i opgaven.

2.2

Sample likelihood bidraget er i vores tilfælde defineret som:

$$f_{Y_i}(y \mid p_1, p_2, p_3) = \ell(p_1, p_2, p_3 \mid y) \quad (3)$$

Antagelsen om uafhængighed sikrer nu, samt ens fordelinger, sikrer, at vi kan skrive Sample Likelihood funktionen som produktet af de individuelle likelihood bidrag, dvs:

$$L(p_1, p_2, p_3 \mid y) = \prod_{i=1}^{573} \ell(p_1, p_2, p_3 \mid y) = p_1^{s_1} \cdot p_2^{s_2} \cdot p_3^{s_3} \quad (4)$$

Hvor det sidste lighedstegn følger af, at vi kan skrive sandsynlighedsfunktionen som vist i (2), og s_i da tæller hvor ofte hver observation forekommer (og et tal opløftet i 0 er 1, hvorfor det kun er den observation, der kommer med).

2.3

Idet parametrene skal summe til en, kan vi udtrykke enhver parameter ved to andre, f.eks: $p_3 = 1 - p_1 - p_2$. Der er således to parametre vi frit kan vælge (og dermed er frie). Altså:

$$\theta = (p_1, p_2) \quad (5)$$

Det skal dog gælde, at de begge er ikke negative og summer til 1 eller derunder. Parameterrummet er altså:

$$\Theta = \{(p_1, p_2) \in \mathbb{R}^2 \mid p_1 \geq 0, p_2 \geq 0, p_1 + p_2 \leq 1\} \quad (6)$$

Bemærk, at vi ligeledes kunne have valgt p_2 som den faste eller p_1 som den faste. Sample likelihood funktionen kan altså skrives som:

$$L_{Y_i}(\theta \mid y) = p_1^{s_1} \cdot p_2^{s_2} \cdot (1 - p_1 - p_2)^{s_3} \quad (7)$$

Mens log likelihood funktionen er givet ved:

$$\log L_{Y_i}(\theta | y) = s_1 \cdot \log(p_1) + s_2 \cdot \log(p_2) + s_3 \cdot \log(1 - p_1 - p_2) \quad (8)$$

2.4

Da vi opskrev sample likelihood funktionen antog vi i.i.d, dvs. at fordelinger kom fra en tæthed (DGP), og at denne er ens for alle stokastiske variable, dvs. P . Samtidig antog vi uafhængighed, da vi skrev den som produktet af de marginale sample likelihood bidrag. Antagelserne virker rimelige, idet der er tale om et tilfældigt udvalgt sample (dvs. uafhængighed er opfyldt), samtidig med, at selve fordelingen blot angiver sandsynligheder for selve udfaldene og dermed også virker rimelig.

2.5

Vi finder nu score vektorerne:

$$\begin{aligned} \frac{\delta L_{Y_i}}{\delta p_1} &= \frac{s_1}{p_1} - \frac{s_3}{1 - p_1 - p_2} \\ \frac{\delta L_{Y_i}}{\delta p_2} &= \frac{s_2}{p_2} - \frac{s_3}{1 - p_1 - p_2} \end{aligned}$$

Vi ser, at:

$$\begin{aligned} \frac{s_1}{p_1} &= \frac{s_2}{p_2} \\ \frac{p_1}{s_1} &= \frac{p_2}{s_2} \\ p_1 &= \frac{p_2}{s_2} \cdot s_1 \end{aligned}$$

Samtidig har vi fra den første ligning, at:

$$\begin{aligned} s_1 \cdot \left(1 - \frac{p_2}{s_2} \cdot s_1 - p_2\right) &= s_3 \cdot \frac{p_2}{s_2} \cdot s_1 \\ s_1 \cdot \frac{s_2 - p_2 s_1 - p_2 s_2}{s_2} &= \frac{s_3 p_2 s_1}{s_2} \\ s_2 - p_2 s_1 - p_2 s_2 &= s_3 p_2 \\ s_2 &= p_2 \cdot (s_1 + s_2 + s_3) \\ p_2 &= \frac{s_2}{s_1 + s_2 + s_3} = \frac{s_2}{n} \end{aligned}$$

Og da:

$$p_1 = \frac{\frac{s_2}{s_1+s_2+s_3}}{s_2} \cdot s_1 = \frac{s_1}{s_1+s_2+s_3} = \frac{s_1}{n}$$

Da har vi:

$$\hat{\theta} = \begin{cases} p_1 = \frac{121}{573} \\ p_2 = \frac{251}{573} \\ p_3 = \frac{201}{573} \end{cases}$$

Som er nøjagtig lig de relative frekvenser, og giver mening, idet estimatet med den højeste likelihood i vores simple model jo blot er sandsynligheden for at de forkommer.

Opgave 3

3.1

Den centrale grænseværdisætning tilsiger, at der for Y_i ukorrelerede, identisk stokastiske variable med samme middelværdi og varians gælder følgende asymptotiske approksimation.

$$\bar{X}_n \stackrel{a}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right) \quad (9)$$

Hvor $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n Y_i$. I vores tilfælde har vi $n = 60$, $\mu = 200$, $\sigma^2 = 100$. Altså:

$$\bar{X}_n \stackrel{a}{\sim} N(200, 1.667) \quad (10)$$

3.2

Vi skal udregne $P(60X_n > 12.500) = 1 - P(60X_n \leq 12.500)$, som er ækvivalent med $1 - P(X_n \leq 208.33)$. Det kan vi gøre ved at benytte os af den standardiserede variabel, da vi ved, at den asymptotisk *normalt normalfordelt*:

$$P \left[\frac{\bar{X}_n - 200}{\sqrt{\frac{100}{60}}} \leq \frac{208.33 - 200}{\sqrt{\frac{100}{60}}} \right]$$

$$P[U_n \leq 6.45]$$

Som praktisk talt er 0, givet, U_n er en standard normalfordeling med middelværdi 0 og varians 1.

Da:

$$P(60X_n > 12.500) = 1 - P(X_n \leq 208.33) \approx \quad (11)$$

3.3

Nej, idet Cauchy fordelingen ikke har nogen middelværdi gælder CLT ikke gælder.

Opgave 4

4.1

For Y_i stokastiske variable, der opfylder i.i.d. betingelserne, er Likelihood funktionen givet ved:

$$L(\mu, \sigma^2 \mid Y_1, \dots, Y_n) = \prod_{i=1}^n \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \mu)^2}{2\sigma^2}} \right] \quad (12)$$

Hvorfor log-likelihood funktionen bliver:

$$\begin{aligned} \log L(\mu, \sigma^2 \mid Y_1, \dots, Y_n) &= \sum_{i=1}^n \left[-\frac{1}{2} \log(2\pi\sigma^2) - \frac{(Y_i - \mu)^2}{2\sigma^2} \right] \\ &= \sum_{i=1}^n \left[-\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma^2) - \frac{(Y_i - \mu)^2}{2\sigma^2} \right] \end{aligned}$$

Score bidragene er:

$$\begin{aligned} \frac{\delta \log L}{\delta \mu} &= - \cdot (-2) \frac{Y_i - \mu}{2\sigma^2} = \frac{Y_i - \mu}{\sigma^2} \\ \frac{\delta \log L}{\delta \sigma} &= -\frac{1}{2\sigma^2} + \frac{(Y_i - \mu)^2}{2\sigma^4} \end{aligned}$$

Hvorfor scoren er:

$$\begin{aligned} S(\theta) &= \sum_{i=1}^n s_i(\theta) \\ &= \begin{bmatrix} -n\mu + \sum_{i=1}^n Y_i \\ -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^n (Y_i - \mu)^2}{2\sigma^4} \end{bmatrix} \end{aligned}$$

Som begge skal være 0 for vores MLE estimat. Fra den første har vi:

$$\begin{aligned} -n\mu + \sum_{i=1}^n Y_i &= 0 \\ \mu &= \frac{1}{n} \sum_{i=1}^n Y_i \end{aligned}$$

Som blot er det aritmetiske gennemsnit. Fra den anden har vi:

$$-n\sigma^2 = -\sum_{i=1}^n (Y_i - \mu)^2$$

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \mu)^2$$

Som blot er lig det andet centrale moment, variansen.

4.2

Vi starter med middelværdien:

$$\mu = \frac{627.6}{125} \approx 5.02$$

Og tager herefter variansen:

$$\sigma^2 \approx \frac{3807.2 + 125 \cdot 5.02^2 - 2 \cdot 5.02 \cdot 627.6}{125} \approx 5.24$$

4.3

Under de nye antagelser bliver likelihood funktionen:

$$L(\phi \mid Y_1, \dots, Y_n) = \prod_{i=1}^n \left[\frac{1}{\sqrt{2\pi\phi^2}} e^{-\frac{(Y_i - \phi)^2}{2\phi^2}} \right] \quad (13)$$

Hvorfor log-likelihood funktionen er:

$$\log L(\phi \mid Y_1, \dots, Y_n) = \sum_{i=1}^n \left[-\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\phi^2) - \frac{(Y_i - \phi)^2}{2\phi^2} \right] \quad (14)$$

For at bestemme score funktionen tages en afledt af gangen. Det er klart, at det første led er 0, da det ikke afhænger af ϕ . Det andet er:

$$\sum_{i=1}^n -\frac{1}{2\phi^2} = -\frac{n}{2\phi^2} \quad (15)$$

Det tredje er:

$$= -\sum_{i=1}^n \left[\frac{-2 \cdot (Y_i - \phi)}{2\phi} - \frac{2(Y_i - \phi)^2}{\phi^2} \right]$$

$$\begin{aligned}
& - \sum_{i=1}^n \left[-\frac{(Y_i - \phi)}{\phi} - \frac{(Y_i - \phi)^2}{2\phi^2} \right] \\
& = - \sum_{i=1}^n \left[-\frac{(Y_i - \phi)}{\phi} - \frac{(Y_i - \phi)^2}{2\phi^2} \right] \\
& = - \sum_{i=1}^n \left[\frac{-2\phi Y_i + 2\phi^2 - Y_i^2 + 2Y_i\phi - \phi^2}{2\phi^2} \right] \\
& = - \sum_{i=1}^n \left[\frac{\phi^2 - Y_i^2}{2\phi^2} \right] \\
& = \frac{\sum_{i=1}^n Y_i^2 - n\phi^2}{2\phi^2}
\end{aligned}$$

Vi kan nu sætte det hele sammen:

$$S(\phi) = \frac{\sum_{i=1}^n Y_i^2 - n\phi^2 - n\phi}{2\phi^2} \quad (16)$$

Og få det ønskede.

4.4

Estimatoren opfylder, at:

$$\begin{aligned}
\sum_{i=1}^n Y_i^2 - n\phi^2 - n\phi &= 0 \\
\phi^2 + \phi - \frac{1}{n} \sum_{i=1}^n Y_i^2 &= 0
\end{aligned}$$

Som er en andengradslikning med $a = 1$, $b = 1$, $c = -\frac{1}{n} \sum x^2$ og har løsningen:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Hvorfor vi får:

$$\hat{\theta} = \frac{-1 \pm \sqrt{\frac{4}{n} \sum x^2 + 1}}{2} \quad (17)$$

Men idet vi kun er interesseret i den positive parameter får vi:

$$\hat{\theta} = \frac{\sqrt{\frac{4}{n} \sum Y_i^2 + 1} - 1}{2} \quad (18)$$

4.5

Estimatet bliver:

$$\hat{\theta} = \frac{\sqrt{\frac{4}{125} \cdot 3807.2 + 1} - 1}{2} \approx 5.041 \quad (19)$$

Som altså ligger nogenlunde mellem middelværdien og variansen for den *rigtige* normalfordeling.

Opgave 5

5.1

Likelihood funktionen er;

$$L(\alpha, \beta \mid y_i) = \prod_{i=1}^n \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot y_i^{\alpha-1} \cdot e^{-\beta y_i} \quad (20)$$

Mens log likelihood funktionen er:

$$\begin{aligned} \log L(\alpha, \beta \mid y_i) &= \sum_{i=1}^n [\alpha \cdot \log(\beta) - \log(\Gamma(\alpha)) + (\alpha - 1) \log(y_i) - \beta y_i] \\ &= n \cdot \alpha \cdot \log(\beta) - n \cdot \log(\Gamma(\alpha)) + (\alpha - 1) \sum_{i=1}^n \log(y_i) - \beta \sum_{i=1}^n y_i \end{aligned}$$

5.2

FOC er, at begge score funktioner skal være lig 0. Vi starter med score funktionen mht. α :

$$\begin{aligned} \frac{\delta L}{\delta \hat{\alpha}} &= n \log(\hat{\beta}) - n \frac{\Gamma(\hat{\alpha})'}{\Gamma(\hat{\alpha})} + \sum_{i=1}^n \log(y_i) = 0 \quad \text{hvor det midterste led følger af kvotientreglen} \\ \frac{\delta L}{\delta \hat{\beta}} &= \frac{n \cdot \hat{\alpha}}{\hat{\beta}} - \sum_{i=1}^n y_i = 0 \end{aligned}$$

5.3

Vi løser ligning 2 for $\frac{\hat{\alpha}}{\hat{\beta}}$ og får:

$$\begin{aligned} \frac{n \cdot \hat{\alpha}}{\hat{\beta}} - \sum_{i=1}^n y_i &= 0 \\ \frac{n \cdot \hat{\alpha}}{\hat{\beta}} &= \sum_{i=1}^n \log(y_i) \end{aligned}$$

$$\frac{\hat{\alpha}}{\hat{\beta}} = \frac{1}{n} \sum_{i=1}^n y_i$$

Resultatet virker meget rimeligt, idet middelværdien af Gammafordelingen netop er $\frac{\hat{\alpha}}{\hat{\beta}}$, som er det, der står på højre siden.

5.4

Fra ovenstående har vi:

$$\hat{\beta} = \frac{n \cdot \hat{\alpha}}{\sum_{i=1}^n y_i}$$

Vi kan da indsætte i den første FOC og få:

$$\begin{aligned} n \log\left(\frac{n \cdot \hat{\alpha}}{\sum_{i=1}^n y_i}\right) - n \frac{\Gamma(\hat{\alpha})'}{\Gamma(\hat{\alpha})} + \sum_{i=1}^n \log(y_i) \\ n \log(n \cdot \hat{\alpha}) - n \cdot \log \sum_{i=1}^n y_i - n \frac{\Gamma(\hat{\alpha})'}{\Gamma(\hat{\alpha})} + \sum_{i=1}^n \log(y_i) \\ n \left[\log(n \cdot \hat{\alpha}) - \log \sum_{i=1}^n y_i - \frac{\Gamma(\hat{\alpha})'}{\Gamma(\hat{\alpha})} \right] + \sum_{i=1}^n \log(y_i) = 0 \end{aligned}$$

Som giver det ønskede resultat.

5.5

Nedenfor er de centrale momenter, samt percentiler for observationerne angivet:

Y					
Percentiles		Smallest			
1%	1.0608	.66647			
5%	1.9581	.99459			
10%	2.46845	1.0608	Obs	250	
25%	3.3655	1.0777	Sum of wgt.	250	
50%	4.47325		Mean	4.81098	
		Largest	Std. dev.	2.120183	
75%	5.9675	11.044			
90%	7.6491	11.352	Variance	4.495175	
95%	8.4153	12.028	Skewness	.9009869	
99%	11.352	12.394	Kurtosis	4.0494	

Figure 2: Beskrivende statistik gamma fordelingen

5.6

Hvis vi får at vide, at $\hat{\alpha} \approx 5.099$ må kan $\hat{\beta}$ findes ved:

$$\hat{\beta} = \frac{250 \cdot 5.099}{1202.75} \approx 1.056 \quad (21)$$