Eksamen på Økonomistudiet vinteren 2016-17
Sandsynlighedsteori og Statistik
2. årsprøve
21. februar, 2017

(3-timers prøve med hjælpemidler)

Rettevejledning

Opgaven består af tre delopgaver, som alle skal besvares. De tre opgaver kan regnes uafhængigt af hinanden. Opgave 1 og 2 indgår tilsammen med samme vægt som opgave 3.

Opgave 1

I denne opgave ses på et marked for brugte cykler. Vi antager, at værdien af brugte cykler kan beskrives ved en normalfordeling. Lad X være den stokastiske variabel, som angiver værdien af den brugte cykel (i kr). Der gælder, at $X \sim N(1000, 8100)$.

1. Andelen af brugte cykler på markedet som har en værdi, der er større end 1100 kr., $P(X > 1100)$ kan beregnes idet man anvender fordelingsfunktionen for normalfordelingen

$$\begin{aligned} P(X > 1100) &= 1 - P(X \leq 1100) \\ &= 1 - P\left(\frac{X - 1000}{\sqrt{8100}} \leq \frac{1100 - 1000}{\sqrt{8100}}\right) \\ &= 1 - \Phi\left(\frac{1100 - 1000}{\sqrt{8100}}\right) \\ &= 1 - \Phi\left(\frac{100}{90}\right) \\ &= 1 - 0.867 \\ &= 0.133. \end{aligned}$$

2. Vi antager nu, at prisen på en brugt cykel afhænger af den sande værdi X, men at det ikke er muligt præcist at vurdere værdien af cyklen. Derfor afhænger prisen også af en "målefejl", Z. Z er en stokastisk variabel, som er normalfordelt, $Z \sim N(0, 100)$ og uafhængig af X. Prisen, Y, er givet ved

$$Y = 50 + X + Z.$$

Fordelingen af Y kan bestemmes som en normalfordeling fordi summen af to normalfordelte variable er normalfordelt. Middelværdien $E(Y)$ og variansen af prisen $Var(Y)$:

$$\begin{aligned} E(Y) &= 50 + E(X) + E(Z) = 50 + 1000 + 0 = 1050. \\ Var(Y) &= Var(X) + Var(Z) = 8100 + 100 = 8200. \end{aligned}$$

I den sidste udregning anvendes at X og Z er uafhængige

3. Kovariansen mellem værdien, X, og prisen, Y, kan beregnes ved at anvende regleregler for varianser og kovarianser

$$\begin{aligned} Cov(X,Y) &= Cov(X, 50 + X + Z) \\ &= Cov(X,X) + Cov(X,Z) \\ &= Var(X) + 0 \\ &= 8100. \end{aligned}$$

   Da $Cov(X,Y) \neq 0$ er X og Y ikke uafhængige.

4. Den forventede pris, når kvaliteten er lig 1100, $E(Y|X=1100)$:

$$\begin{aligned} E(Y|X=1100) &= E(Y) + \frac{Cov(X,Y)}{Var(X)}(1100 - E(X)) \\ &= 1050 + \frac{8100}{8100}(1100 - 1000) \\ &= 1050 + 100 = 1150. \end{aligned}$$

Opgave 2

Lad X være ligefordelt på $(1,2)$ dvs. X har tæthed $p(x) = \mathbf{1}(1 < x < 2)$.

1. Find $E(X)$ og $Var(X)$.

   Facit: $E(X) = \int_1^2 x dx = \frac{1}{2}(4-1) = \frac{3}{2}$ og $Var(X) = E(X^2) - (EX)^2$ så

$$E(X^2) = \int_1^2 x^2 dx = \frac{1}{3}(8-1) = \frac{7}{3},$$

   dvs

$$Var(X) = \frac{7}{3} - \left(\frac{3}{2}\right)^2 = \frac{7}{3} - \frac{9}{4} = \frac{28-27}{12} = \frac{1}{12}.$$

2. Vi sætter nu $Y = X - 1$. Find $E(Y)$ og $Var(Y)$.

   Facit: $E(Y) = E(X) - 1 = \frac{1}{2}$ og $Var(Y) = Var(X)$.

3. Find tætheden $q(y)$ for Y.

Facit: Enten ses direkte $U(0,1)$ eller ved $t(x) = x - 1$ så $t'(x) = 1$ og

$$\begin{aligned} q(y) &= p\left(t^{-1}(y)\right) = p(y+1) \\ &= \mathbf{1}(1 < y+1 < 2) \\ &= \mathbf{1}(0 < y < 1). \end{aligned}$$

4. Find $E\left(X|X > \frac{3}{2}\right)$.

   Facit: Enten ses direkte som $\frac{3}{2} + \left(2 - \frac{3}{2}\right)/2 = \frac{3}{2} + \frac{1}{4} = \frac{7}{4}$ eller

$$P\left(X > \frac{3}{2}\right) = \int_{\frac{3}{2}}^{2} dx = 2 - \frac{3}{2} = \frac{1}{2},$$

   så

$$\begin{aligned} E\left(X|X > \frac{3}{2}\right) &= \frac{\int_{\frac{3}{2}}^{2} x dx}{\frac{1}{2}} \\ &= 2\frac{1}{2}\left(4 - \frac{9}{4}\right) \\ &= \frac{16}{4} - \frac{9}{4} = \frac{7}{4}. \end{aligned}$$

Opgave 3

1. Til den beskrivende statistik bemærkes det fx at den empiriske fordeling kun har positiv støtte (her kun $y_i > 1000\$$), at den efter udregnet skewness er klart højreskæv, og at den efter udregnet kurtosis har markant tykkere haler end normalfordelingen.

2. For at vise at $F_{Y_i}(y \mid \theta)$ er fordelingsfunktionen, kan man differentiere og sammenligne med tæthedfunktionen:

$$\frac{\partial}{\partial y} F_{Y_i}(y \mid \theta) = \frac{\partial}{\partial y}\left(1 - c^{\theta} y^{-\theta}\right) = \theta c^{\theta} y^{-\theta-1} = f_{Y_i}(y \mid \theta).$$

3. Med den valgte fordelingsantagelse er likelihood bidraget givet ved

$$\ell(\theta \mid y_i) = f_{Y_i}(y_i \mid \theta) = \theta c^{\theta} y_i^{-\theta-1},$$

   som er antaget identisk for alle $i = 1, 2, ..., n$. Så er log-likelihood bidraget givet som

$$\log \ell(\theta \mid y_i) = \log(\theta) + \theta \log(c) - (\theta + 1)\log(y_i).$$

Under antagelse af uafhængighed er den samlede likelihood funktion givet som

$$L_n(\theta) = \prod_{i=1}^{n} \ell(\theta \mid y_i) = \prod_{i=1}^{n} \theta c^{\theta} y_i^{-\theta-1},$$

sådan at

$$\log L_n(\theta) = \sum_{i=1}^{n} \{\log(\theta) + \theta \log(c) - (\theta+1)\log(y_i)\}.$$

4. Score-bidraget fra observation y_i er givet som

$$s_i(\theta) = \frac{\partial \log \ell(\theta \mid y_i)}{\partial \theta} = \frac{1}{\theta} + \log(c) - \log(y_i) = \frac{1}{\theta} - \log\left(\frac{y_i}{c}\right),$$

så scoren er

$$S(\theta) = \sum_{i=1}^{n} s_i(\theta) = \sum_{i=1}^{n} \left\{\frac{1}{\theta} - \log\left(\frac{y_i}{c}\right)\right\}.$$

Dermed er første-ordens betingelsen givet ved

$$S(\hat{\theta}) = \sum_{i=1}^{n} \left\{\frac{1}{\hat{\theta}} - \log\left(\frac{y_i}{c}\right)\right\} = \frac{n}{\hat{\theta}} - \sum_{i=1}^{n} \log\left(\frac{y_i}{c}\right) = 0,$$

som løses hvor

$$\begin{aligned} \frac{n}{\hat{\theta}_n} &= \sum_{i=1}^{n} \log\left(\frac{y_i}{c}\right) \\ \hat{\theta}_n &= \frac{n}{\sum_{i=1}^{n} \log\left(\frac{y_i}{c}\right)}. \end{aligned}$$

5. Hale-sandsynligderne udregnes som

$$P(Y_i > y) = 1 - P(Y_i \leq y) = 1 - (1 - c^{\theta} y^{-\theta}) = c^{\theta} y^{-\theta}.$$

Baseret på estimatet $\hat{\theta}_n = 2.64$ og $c = 1000$ findes

$$\begin{aligned} P(Y_i &> 2500) = 1000^{2.64} 2500^{-2.64} = 0.089 \\ P(Y_i &> 3000) = 1000^{2.64} 3000^{-2.64} = 0.055. \end{aligned}$$

Model-kontrol kan baseres på en sammenligning mellem den antagede Pareto-fordeling med $\hat{\theta}_n$ indsat og den empiriske fordeling af $\{y_i\}_{i=1}^{n}$. For det konkrete tilfælde passer de udregnede sandsynligheder udmærket med de rapporterede fraktiler i tabellen.

6. Hesse-bidraget fra observation y_i er givet som

$$H_i(\theta) = \frac{\partial^2 \log \ell(\theta \mid Y_i)}{\partial\theta\partial\theta} = \frac{\partial s_i(\theta)}{\partial\theta} = \frac{\partial}{\partial\theta}\left(\frac{1}{\theta} - \log\left(\frac{Y_i}{c}\right)\right) = \frac{-1}{\theta^2}.$$

Så er informationen, evalueret i den sande værdi, θ_0, givet som

$$I(\theta_0) = E(-H_i(\theta_0)) = \theta_0^{-2}.$$

Variansen på estimatoren er derfor

$$V(\hat{\theta}_n) = \frac{1}{n} I(\theta_0)^{-1} = \frac{\theta_0^2}{400}.$$

Med estimatoren indsat i stedet for θ_0 fås

$$V(\hat{\theta}) \approx \frac{\hat{\theta}_n^2}{400} = \frac{2.64^2}{400} = 0.0174,$$

og derfor er $\mathsf{se}(\hat{\theta}_n) = \sqrt{0.0174} = 0.132$.

7. Under antagelserne for den centrale grænseværdisætning, gælder at

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left(0, I(\theta_0)^{-1}\right).$$

Med informationen $I(\hat{\theta}_n)$ fås den asymptotiske approximation

$$\hat{\theta}_n \overset{a}{\sim} N\left(\theta_0, 0.0174\right).$$

For at finde et 95% konfidens interval bruges fraktilerne i normalfordelingen, så der er 95% sandsynlighed for at den sande værdi, θ_0, ligger i intervallet

$$\begin{aligned}\{\underline{\theta} \leq \theta_0 \leq \overline{\theta}\} &= \{\hat{\theta}_n - 1.96 \cdot \mathsf{se}(\hat{\theta}_n) \leq \theta_0 \leq \hat{\theta}_n + 1.96 \cdot \mathsf{se}(\hat{\theta}_n)\} \\ &= \{2.38 \leq \theta_0 \leq 2.90\}.\end{aligned}$$