Eksamen på Økonomistudiet vinteren 2017-18
Sandsynlighedsteori og Statistik
2. årsprøve
14. februar, 2018
(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

Antag at indkomsten X for en passende indtægtsgruppe er ligefordelt på intervallet [1,2] (angivet i 100000 DKK).

1. Opskriv tætheden $p(x)$ for X og find $E(X)$.

   Facit: $p(x) = 1(1 < x < 2)$ og $E(X) = 1.5$.

   Man kan se det direkte, eller:

   $$\int xp(x)\,dx = \left[\frac{1}{2}x^2\right]_1^2 = 1.5.$$

2. Man ønsker at se på fordelingen af $Y = \log(X)$. Find tætheden $q(y)$ for Y.

   Facit:

   $$\begin{aligned} q(y) &= p\left(t^{-1}(y)\right)\left|\partial t^{-1}(y)/\partial y\right| \\ &= 1(1 < \exp(y) < 2)\exp(y) \\ &= \exp(y)\,1(0 < y < \log(2)). \end{aligned}$$

3. Vis at $E(Y) = 2\log(2) - 1$.

   Hint: Der gælder at $\int \exp(x)\,xdx = \exp(x)(x-1)$ og tilsvarende at $\int \log(x)\,dx = x(\log(x) - 1)$.

   Facit: Enten direkte:

   $$\begin{aligned} E(Y) &= E\log(X) \\ &= \int \log(x)\,p(x)\,dx \\ &= \int_1^2 \log(x)\,dx \\ &= \left[x(\log(x) - 1)\right]_1^2 \\ &= 2(\ln(2) - 1) + 1 \\ &= 2\log(2) - 1. \end{aligned}$$

Eller:

$$\begin{aligned} E(Y) &= \int_0^{\log 2} y \exp(y)\, dy \\ &= \left[\exp(x)(x-1)\right]_0^{\log(2)} \\ &= 2(\log(2)-1)+1 \\ &= 2\log(2)-1. \end{aligned}$$

Opgave 2

I denne opgave undersøges ventetiden (målt i uger) for en arbejdsløs, til vedkommende kommer i arbejde. Ventetiden er defineret som 0 uger, hvis personen kommer i arbejde i uge 1, 1 uge hvis personen kommer i arbejde i uge 2 osv. Vi antager, at ventetiden kan beskrives som en stokastisk variabel T, hvor T er geometrisk fordelt med parametren $\theta = 0.2$.

1. Sandsynligheden for at ventetiden er to uger (altså at personen kommer i arbejde i uge 3), $P(T = 2)$ kan beregnes ved brug af sandsynlighedsfunktionen for den geometriske fordeling:

   $$P(T = 2) = (1-\theta)^2 \cdot \theta = 0.8^2 \cdot 0.2 = 0.128$$

   Sandsynligheden for at ventetiden er under to uger $P(T < 2)$, kan beregnes (ved at bruge sandsynlighedsfunktionen for den geometriske fordeling) som

   $$P(T < 2) = P(T = 0) + P(T = 1) = \theta + \theta(1-\theta) = 0.2 + 0.2 \cdot 0.8 = 0.36$$

2. Den forventede ventetid $E(T)$ kan beregnes som

   $$E(T) = \frac{1-\theta}{\theta} = \frac{0.8}{0.2} = 4$$

3. Vi antager nu, at der er 100 arbejdsløse tilknyttet et jobcenter. Hvis den arbejdsløse ikke selv finder arbejde i løbet af de to første uger, skal der laves en jobplan. Denne opgave kan løses ved at definere en ny stokastisk variabel, som angiver, om der skal laves en jobplan for person i

   $$\begin{aligned} Y_i &= 1 \text{ hvis } T_i \geq 2 \\ Y_i &= 0 \text{ hvis } T_i < 2. \end{aligned}$$

Y_i er Bernoulli-fordelte stokastiske variable, og der gælder at

$$E(Y_i) = P(Y_i = 1) = P(T_i \geq 2) = 1 - P(T_i < 2) = 1 - 0.36 = 0.64$$

Det forventede antal arbejdsløse, hvor jobcentret skal lave en jobplan, kan så beregnes som

$$E(Y_1 + Y_2 + ... + Y_{100}) = E(Y_1) + E(Y_2) + ... + E(Y_{100}) = 100 \cdot 0.64 = 64.$$

Vi antager nu, at der er to arbejdsløse, og begge deres ventetider T_1 og T_2 er geometrisk fordelt med parametren $\theta = 0.2$. Desuden antages, at ventetiderne er stokastisk uafhængige.

4. Sandsynligheden for at den samlede ventetid for de to arbejdsløse er 4 uger, $P(T_1 + T_2 = 4)$, kan beregnes på følgende måde

$$\begin{aligned} P(T_1 + T_2 = 4) &= \sum_{t_1=0}^{4} P(T_1 = t_1, T_2 = 4 - t_1) \\ &= \sum_{t_1=0}^{4} P(T_1 = t_1) P(T_2 = 4 - t_1) \\ &= \sum_{t_1=0}^{4} (1-\theta)^{t_1} \cdot \theta (1-\theta)^{4-t_1} \cdot \theta \\ &= 5(1-\theta)^4 \cdot \theta^2 \\ &= 0.082. \end{aligned}$$

Opgave 3

1. For identisk Erlang-fordelte stokastiske variable er sample likelihood bidraget givet som

$$\ell(\lambda \mid y_i) = f_{Y_i}(y_i \mid \lambda) = \lambda^2 y_i \exp(-\lambda y_i),$$

så

$$\log \ell(\lambda \mid y_i) = 2\log(\lambda) + \log(y_i) - \lambda y_i.$$

Under uafhængighed fås de tilsvarende for hele datasættet:

$$L_n(\lambda) = \prod_{i=1}^{n} \ell\left(\lambda \mid y_i\right) = \prod_{i=1}^{n} \lambda^2 y_i \exp\left(-\lambda y_i\right)$$

og

$$\log L_n(\lambda) = \sum_{i=1}^{n} \log \ell\left(\lambda \mid y_i\right) = \sum_{i=1}^{n} \left\{2\log(\lambda) + \log(y_i) - \lambda y_i\right\}.$$

2. Scorebidraget er givet ved

$$s_i(\lambda) = \frac{\partial \log \ell\left(\lambda \mid y_i\right)}{\partial \lambda} = \frac{2}{\lambda} - y_i,$$

så scoren er

$$S(\lambda) = \sum_{i=1}^{n} \left\{\frac{2}{\lambda} - y_i\right\} = \frac{2n}{\lambda} - \sum_{i=1}^{n} y_i.$$

Førsteordensbetingelsen er derfor

$$S(\hat{\lambda}_n) = 0,$$

sådan at $\hat{\lambda}_n$ er bestemt som

$$\begin{aligned} \frac{2n}{\hat{\lambda}_n} - \sum_{i=1}^{n} y_i &= 0 \\ \hat{\lambda}_n &= \frac{2n}{\sum_{i=1}^{n} y_i}. \end{aligned}$$

Fra den beskrivende statistik er gennemsnittet $m_y = 14.334$, sadan at $\sum_{i=1}^{n} y_i = 740 \cdot 14.334 = 10607.16$ og

$$\hat{\lambda}_n = \frac{2 \cdot 740}{10607.16} = 0.13953.$$

3. Bidraget til Hesse-matricen er defineret som

$$\begin{aligned} H_i(\lambda) &= \frac{\partial^2 \log \ell\left(\lambda \mid Y_i\right)}{\partial \lambda \partial \lambda} \\ &= \frac{\partial}{\partial \lambda} s_i(\theta) \\ &= \frac{\partial}{\partial \lambda}\left(\frac{2}{\lambda} - Y_i\right) \\ &= -\frac{2}{\lambda^2}. \end{aligned}$$

Dermed er informationen,

$$I(\lambda_0) = E\left(-H_i(\lambda_0)\right) = \frac{2}{\lambda_0^2}.$$

Det følger, at variansen på estimatoren er

$$V(\hat{\lambda}_n) = \frac{I(\lambda_0)^{-1}}{n} = \frac{\lambda_0^2}{2n},$$

som kan approksimeres med

$$V(\hat{\lambda}_n) \approx \frac{\hat{\lambda}_n^2}{2n} = \frac{0.13953^2}{2 \cdot 740} = 0.000013154.$$

Dermed er $\mathsf{se}(\hat{\lambda}_n) = \sqrt{V(\hat{\lambda}_n)} = \sqrt{\frac{0.13953^2}{2 \cdot 740}} = 0.0036269$.

4. Som kontrol af antagelsen om en Erlang fordeling udregnes de teoretiske momenter for Erlang($\hat{\lambda}_n$) og sammenlignes med resultaterne fra den beskrivende statistik, der fås

	Ventetid		Teoretisk
Gennemsnit	14.334	$2/\hat{\lambda}_n$	14.334
Standardafvigelse	9.832	$\sqrt{2/\hat{\lambda}_n^2}$	10.136
Skewness	1.131	$\sqrt{2}$	1.414
Kurtosis	4.167		6

Overordner er der rimelig overensstemmelse med modelantagelsen, omend Erlang fordelingen teoretisk skulle have lidt større skewness og kurtosis, svarende til et relativt større antal lange ventider.

5. En gennemsnitlig ventetid på 15 minutter svarer til at $\lambda_0 = 2/15 = 0.13333$. For at teste dette anvendes hypoteser

$$H_0 : \lambda_0 = 2/15 \quad \text{mod} \quad H_A : \lambda_0 \neq 2/15.$$

Teststørrelsen er givet som

$$z_n(\lambda_0 = 2/15) = \frac{\hat{\lambda}_n - 2/15}{\mathsf{se}(\hat{\lambda}_n)} = \frac{0.13953 - 2/15}{0.0036269} = 1.7085.$$

Hvis nul-hypotesen er korrekt, er størrelsen fordelt som en standard normal-fordeling, $N(0,1)$, så på $\alpha = 0.05$ signifikans-niveau er den kritiske værdi 1.96. Hypotesen om en gennemsnitlig ventetid på 15 minutter kan derfor ikke afvises, selv om det er ret tæt på. Man kan udregne p-værdien for dette tilfælde som

$$p_n(\lambda_0 = 2/15) = 2\left(1 - \Phi(1.7085)\right) = 0.08754.$$

6. Den udvidede model kan formuleres om en betinget model for $Y_i \mid x_i$, hvor parameteren i den betingede Erlang fordeling nu er en funktion af x_i:

$$Y_i \mid x_i \stackrel{d}{=} \text{Erlang}(\lambda_i) \quad \text{med} \quad \lambda_i = \beta + \delta x_i,$$

hvor x_i er indikatorvariablen for en helligdag. Så er likelihood bidraget givet ved

$$\begin{aligned} \ell\left(\beta, \delta \mid y_i, x_i\right) &= f_{Y_i|X_i}\left(y_i \mid x_i; \beta, \delta\right) \\ &= \left(\beta + \delta x_i\right)^2 y_i \exp\left(-\left(\beta + \delta x_i\right) y_i\right), \end{aligned}$$

og likelihood funktionen for hele datasættet er

$$\begin{aligned} L_n(\beta, \delta) &= \prod_{i=1}^{n} \ell\left(\beta, \delta \mid y_i, x_i\right) \\ &= \prod_{i=1}^{n} \left(\beta + \delta x_i\right)^2 y_i \exp\left(-\left(\beta + \delta x_i\right) y_i\right). \end{aligned}$$