Eksamen på Økonomistudiet sommeren 2019

Sandsynlighedsteori og Statistik

2. årsprøve

28. August, 2019

(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

1. Den marginale sandsynlighedsfunktion $p(x)$ kan findes ved $p(1) = P(X = 1) = P(X = 1, Y = 1) + P(X = 1, Y = 2) = 0.3$ og $p(2) = P(X = 2) = 1 - p(1) = 0.7$.
   Vi har dermed sandsynlighedsfunktionen

x	1	2
$p(x)$	0.3	0.7

2. $\mathbb{E}(X^2) = P(X = 1) \cdot 1^2 + P(X = 2) \cdot 2^2 = 0.3 + 0.7 \cdot 4 = 3.1$.

3. Til at beregne variansen bruger vi, at $\mathbb{E}(X) = P(X = 1) \cdot 1 + P(X = 2) \cdot 2 = 0.3 + 0.7 \cdot 2 = 1.7$ sammen med svaret ovenover. Vi har

$$\begin{aligned} Var(X) &= \mathbb{E}(X^2) - \mathbb{E}(X)^2 \\ &= 3.1 - 1.7^2 \\ &= 0.21 \end{aligned}$$

   .

4. Vi bruger regneregler for varians

$$\begin{aligned} Var(Z) &= Var(-2 \cdot X + 3) \\ &= (-2)^2 Var(X) \\ &= 4 \cdot 0.21 \\ &= 0.84 \end{aligned}$$

   Til at beregne korrelationen mellem X og Z beregner vi først kovariansen ved at at

bruge regneregler herfor:

$$\begin{aligned} Cov(X, Z) &= Cov(X, -2 \cdot X + 3) \\ &= -2Cov(X, X) \\ &= -2Var(X) \\ &= -0.42 \end{aligned}$$

Vi indsætter i formlen for korrelationen

$$\begin{aligned} corr(X, Z) &= \frac{Cov(X, Z)}{\sqrt{Var(X)}\sqrt{Var(Z)}} \\ &= \frac{-0.42}{\sqrt{0.21}\sqrt{0.84}} \\ &= -1 \end{aligned}$$

Opgave 2

1. Fordelingsfunktionen for en Eksponential-fordelt stokastisk variabel er

$$\begin{aligned} F(x) = P(X \leq x) &= \int_0^x \frac{1}{2} \exp(-y/2) dy \\ &= [-\exp(y/2)]_0^x \\ &= 1 - \exp(x/2) \end{aligned}$$

2. Vi finder vha. Fordelingsfunktionen

$$\begin{aligned} P(X \geq 0.5) &= 1 - P(X \leq 0.5) + \underbrace{P(X = 0.5)}_{=0} \\ &= 1 - F(0.5) \\ &= 1 - [1 - \exp(-0.25)] \\ &\approx 0.7788 \end{aligned}$$

3. Vi har, at $Y = t(X) = \log(X)$ sådan at grænserne for den nye stokastiske variabel er

$$\begin{aligned} v &= \inf_{x \geq 0} t(x) = -\infty \\ h &= \sup_{x \geq 0} t(x) = \infty \end{aligned}$$

og Y derfor er fordelt på intervallet $(-\infty, \infty) = \mathbb{R}$.

4. Vi har at

$$t^{-1}(y) = \exp(y)$$
$$\frac{\partial t^{-1}(y)}{\partial y} = \exp(y).$$

Vi får derfor tæthedsfunktionenen for $Y \in \mathbb{R}$:

$$\begin{aligned} q(y) &= p(t^{-1}(y)) \left| \frac{\partial t^{-1}(y)}{\partial y} \right| \\ &= \frac{1}{2} \exp(-\exp(y)/2) \exp(y) \\ &= \frac{1}{2} \exp(y - \exp(y)/2), y \in (-\infty, \infty) \end{aligned}$$

Opgave 3

1. Parameterrummet er $\Theta = \{\theta : -\infty < \theta < \infty\} = \mathbb{R}$ da dette sikrer, at skala-parametren, $\exp(\theta) > 0$.

2. Vi har, at likelihood bidragene for hver klasse er $\ell(\theta|y_i) = p(y_i) = \exp(-\exp(\theta))\frac{\exp(\theta \cdot y_i)}{y_i!}$ og log-likelihood bidraget er $\log(\ell(\theta|y_i)) = -\exp(\theta) + \theta \cdot y_i - \log(y_i!)$. Log-likelihood funktionen bliver således, idet vi udnytter at klasserne er uafhængige,

$$\begin{aligned} \log L_n(\theta) = \log L(\theta|y_1, \ldots, y_n) &= \sum_{i=1}^{n} [-\exp(\theta) + \theta \cdot y_i - \log(y_i!)] \\ &= -n\exp(\theta) + \theta \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \log(y_i!) \end{aligned}$$

3. Første ordens betingelsen (FOC) for den givne model er, at scoren skal være nul,

$$S(\hat{\theta}_n) = \left. \frac{\partial \log L_n(\theta)}{\partial \theta} \right|_{\theta=\hat{\theta}_n} = 0$$

hvor

$$\begin{aligned}\frac{\partial \log L_n(\theta)}{\partial \theta} &= \sum_{i=1}^{n} \frac{\partial \log(\ell(\theta|Y_i))}{\partial \theta} \\ &= \sum_{i=1}^{n} s_i(\theta) \\ &= \sum_{i=1}^{n} [Y_i - \exp(\theta)] \\ &= (\sum_{i=1}^{n} Y_i) - n\exp(\theta)\end{aligned}$$

sådan at **estimatoren** $\hat{\theta}(Y_1, \ldots, Y_n)$ kan udledes til at være

$$\begin{aligned} S(\hat{\theta}) &= 0 \\ &\Updownarrow \\ n\exp(\theta) &= \sum_{i=1}^{n} Y_i \\ &\Updownarrow \\ \hat{\theta} &= \log\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right). \end{aligned}$$

Ved at indsætte informationen givet i opgaven får vi **estimatet**

$$\begin{aligned}\hat{\theta}_n = \hat{\theta}(y_1, \ldots, y_{75}) &= \log\left(\frac{1}{75}\sum_{i=1}^{75} y_i\right) \\ &= \log\left(\frac{1599}{75}\right) \\ &\approx 3.0596.\end{aligned}$$

4. Hesse-matricen er en skalar i dette tilfælde og givet ved den anden-afledte.

$$H_i(\theta) = \frac{\partial^2 \log(\ell(\theta|y_i))}{\partial^2 \theta} = \frac{\partial s_i(\theta)}{\partial \theta} = -\exp(\theta)$$

og dermed er informationen

$$\begin{aligned} I(\theta_0) &= \mathbb{E}(-H_i(\theta_0)) \\ &= \mathbb{E}\Big(\exp(\theta_0) \Big) \\ &= \exp(\theta_0) \end{aligned}$$

Ved at indsætte vores estimat fås

$$\begin{aligned} I(\hat{\theta}_n) &= \exp(\hat{\theta}_n) \\ &= \exp(3.0596) \\ &\approx 21.32 \end{aligned}$$

således at variansen

$$\begin{aligned} Var(\hat{\theta}) &= \frac{1}{n} I(\theta_0)^{-1} \\ &= \frac{\exp(-\theta_0)}{n} \end{aligned}$$

kan approksimeres som

$$\begin{aligned} Var(\hat{\theta}) &\approx \frac{1}{n} I(\hat{\theta}_n)^{-1} \\ &= \frac{1}{75 \cdot 21.32} \\ &= .00063 \end{aligned}$$

Det ses at standardafvigelsen bliver $se(\hat{\theta}) = \sqrt{Var(\hat{\theta})} = \sqrt{0.00063} \approx 0.0251$.

5. Vi kan beregne

$$\begin{aligned} P(Y = 22) &= p(22) \\ &= \exp(-\exp(\hat{\theta}_n)) \frac{\exp(\hat{\theta}_n \cdot 22)}{22!} \\ &= \exp(-\exp(3.0596)) \frac{\exp(3.0596 \cdot 22)}{22!} \\ &\approx 0.08383 \end{aligned}$$

dvs. der er ca. 8.4% sandsynlighed for, at der er 22 fraværstimer i en tilfældig klasse.

6. Vi skal teste om $\mathbb{E}(Y_i) = 22$. Vi ved at $\mathbb{E}(Y_i) = \exp(\theta)$ så det svarer til restriktionen $\theta_0 = \log(22) \approx 3.091$ og vi kan opstille vores nul-hypotese

$$\mathcal{H}_0 : \theta_0 = 3.091$$

og alternativ hypotese

$$\mathcal{H}_A : \theta_0 \neq 3.091.$$

Vi beregner vores z-statistik som

$$z_n(\theta_0 = 3.091) = \frac{\hat{\theta}_n - 3.091}{se(\hat{\theta})} = \frac{3.0596 - 3.091}{0.0251} \approx -1.251.$$

Vi ved at $z_n(\theta_0 = 3.091) \overset{a}{\sim} \mathcal{N}(0,1)$ under $\mathcal{H}_0$. Vi kan derfor beregne den kristiske værdi på et 5% signifikans-niveau, $\alpha = 0.05$, som $c = \Phi^{-1}(0.975) \approx 1.96$ (to-sidet test). Da $|z_n| < c$ kan vi IKKE afvise på et 5% signifikansniveau, at det forventede antal fraværstimer i en tilfældig klasse er 22. (p-værdien er $2 \cdot (1 - \Phi(1.251)) \approx 0.2109$, hvilket er noget højere end de 5%)

7. Den betingede model har log-likelihood funktionen

$$\begin{aligned}\log L_n(\theta, \delta) &= \log L(\theta, \delta | y_1, \ldots, y_{75}, x_1, \ldots, x_{75}) \\ &= \sum_{i=1}^{75} \{-\exp(\theta + \delta x_i) + (\theta + \delta x_i) \cdot y_i - \log(y_i!)\}\end{aligned}$$

8. Vi får nu givet, at $L_u = -213.10$ og $L_r = -216.06$. Vi opstiller vores nul-hypotese

$$\mathcal{H}_0 : \delta_0 = 0$$

og alternativ hypotese

$$\mathcal{H}_A : \delta_0 \neq 0.$$

Vi kan beregne vores Likelihood Ratio (LR) test-størrelse som

$$LR(\delta_0 = 0) = 2 \cdot (L_u - L_r) = 2 \cdot (-213.10 + 216.06) = 5.92.$$

Vi ved, at under $\mathcal{H}_0$ er teststørrelsen asymptotisk χ^2 fordelt med 1 frihedsgrad, $LR(\delta_0 = 0) \overset{a}{\sim} \chi^2_1$. Så vi kan beregne den kristiske værdi på et 5% signifikans-niveau, $\alpha = 0.05$, som $c = (\chi^2_1)^{-1}(0.95) \approx 3.84$. Da $LR > 3.84$ kan vi afvise på et 5% signifikansniveau, at $\delta_0 = 0$.

Vi finder altså, at vi på et 5% signifikans niveau kan *afvise*, at der *ikke er forskel* på de gennemsnitlige antal fraværstimer på tværs af landet. Estimationsresultaterne tyder på, at der er signifikant mindre fravær i Jylland.