Eksamen på Økonomistudiet vinteren 2019-20

Sandsynlighedsteori og Statistik

2. årsprøve

4. januar, 2020

(3-timers prøve med hjælpemidler)

Rettevejledning

Opgave 1

1. Sandsynligheden for at Bent ankommer til den manglende lift på tur i er $P(X_i = 1) = p = \frac{1}{90}$.

2. Lad $Y = \sum_{i=1}^{25} X_i$ angive antallet af gange, Bent har mødt den manglende lift nummer 43 ud af sine 25 uafhængige ture. Y er således Binomial-fordelt, $Y \sim Bin(25, \frac{1}{90})$. Sandsynligheden for, at Bent møder den manglende lift 3 gange ud af sine 25 ture er

$$
\begin{aligned}
P(Y=3) &= \begin{pmatrix} 25 \\ 3 \end{pmatrix} \frac{1}{90^3} \left(\frac{89}{90}\right)^{22} \\
&= \frac{25!}{22! \cdot 3!} \frac{1}{90^3} \left(\frac{89}{90}\right)^{22} \\
&\approx 0.0025.
\end{aligned}
$$

3. Sandsynligheden for, at Bent møder den manglende lift mindst 1 gang på sine 25 ture er

$$
\begin{aligned}
P(Y \geq 1) &= 1 - P(Y < 1) \\
&= 1 - P(Y = 0) \\
&= 1 - 0.7563 \\
&= 0.2437
\end{aligned}
$$

   da $P(Y = 0) = \left(\frac{89}{90}\right)^{25} \approx 0.7563$.

4. Vi har nu, at der er to manglende lifter på de sidste 10 ture. Vi lader antallet af gange, Bent møder den manglende lift 2 være angivet ved Z. Vi har at Z også er binomial-fordelt med $Z \sim Bin(10, \frac{1}{90})$. Vi ved, at

$$
\begin{aligned}
\mathbb{E}[Y] &= 25 \cdot \frac{1}{90} = \frac{5}{18} \\
\mathbb{E}[Z] &= 10 \cdot \frac{1}{90} = \frac{1}{9}
\end{aligned}
$$

   sådan at middelværdien af det samlede antal ture, $M = Y + Z$, hvor Bent har mødt en manglende lift, er

$$
\mathbb{E}[M] = \mathbb{E}[Y] + \mathbb{E}[Z] = \frac{35}{90} = \frac{7}{18} \approx 0.3889
$$

(Bemærk, da Y og Z IKKE er uafhængige, kan vi ikke sige, at $M \sim Bin(35, \frac{1}{90})$)

Opgave 2

1. Middelværdien er

$$\begin{aligned}
\mathbb{E}[Y] &= \int_1^5 (6x+4)p(x)dx \\
&= \int_1^5 (6x+4)\frac{1}{4}dx \\
&= [\frac{3}{4}x^2 + x]_1^5 \\
&= 75/4 + 5 - (3/4 + 1) \\
&= 72/4 + 4 \\
&= 22
\end{aligned}$$

2. Denne opgave kan løses som ovenover ved at beregne $\mathbb{E}[Y^2] = \int_1^5 (6x+4)^2 p(x)dx$ og bruge dette i $Var(Y) = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2$. Alternativt, så kan svaret findes ved at finde variansen af X og så bruge regneregler for varians:

$$\mathbb{E}[X] = 3 \text{ (midtpunkt i uniform)}$$

$$\begin{aligned}
\mathbb{E}[X^2] &= \int_1^5 x^2 p(x)dx \\
&= \frac{1}{4}\int_1^5 x^2 dx \\
&= \frac{1}{4}[\frac{1}{3}x^3]_1^5 \\
&= \frac{1}{4}[\frac{125}{3} - \frac{1}{3}] \\
&= \frac{124}{12} \\
&= \frac{31}{3} \approx 41.3333
\end{aligned}$$

og vi får

$$\begin{aligned}
Var(X) &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\
&= \frac{31}{3} - 9 \\
&= \frac{31-27}{3} \\
&= \frac{4}{3}
\end{aligned}$$

Dermed bliver

$$\begin{aligned}
Var(Y) &= Var(6X+4) \\
&= 36Var(X) \\
&= 36\frac{4}{3} \\
&= 48
\end{aligned}$$

3. Vi har transformationen $t(x) = \log(6x+4)$ og
   (a) $p(x) = \frac{1}{4}\mathbf{1}(x \in [1,5]),\ x \in [1,5]$
   (b) grænserne $v = \log(10) \approx 2.30$ og $h = \log(34) \approx 3.53$
   (c) $t^{-1}(z) = (e^z - 4)/6$
   (d) $\frac{\partial t^{-1}(z)}{\partial z} = \frac{1}{6}e^z$
   således at vi har for $z \in (2.30, 3.53)$ at $q(z) = \frac{1}{4}\mathbf{1}((e^z-4)/6 \in [1,5])\frac{1}{6}e^z$, hvor indikator-funktionen altid er 1 for $z \in (2.30, 3.53)$. Vi får altså

$$q(z) = \begin{cases} \frac{1}{24}e^z & \text{hvis } z \in (2.30, 3.53) \\ 0 & \text{hvis } z \notin (2.30, 3.53) \end{cases}$$

Opgave 3

1. Vi har, at likelihood bidraget for hver borger er $\ell(\theta|x_i) = p(x_i) = \frac{\theta^{2x_i}}{x_i!}\exp(-\theta^2)$ og log-likelihood bidraget er $\log(\ell(\theta|x_i)) = 2x_i\log(\theta) - \theta^2 - \log(x_i!)$. Log-likelihood funktionen

bliver, grundet uafhænighed mellem borgerne, således

$$
\begin{aligned}
\log L_n(\theta) = \log L(\theta|x_1, \ldots, x_{251}) &= \sum_{i=1}^{n} \log(\ell(\theta|x_i)) \\
&= \sum_{i=1}^{n} [2x_i \log(\theta) - \theta^2 - \log(x_i!)] \\
&= -n \cdot \theta^2 + 2\log(\theta) \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \log(x_i!)
\end{aligned}
$$

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

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

hvor vi her har at

$$
\begin{aligned}
\frac{\partial \log L_n(\theta)}{\partial \theta} &= \sum_{i=1}^{n} \frac{\partial \log(\ell(\theta|X_i))}{\partial \theta} \\
&= \sum_{i=1}^{n} s_i(\theta) \\
&= \sum_{i=1}^{n} \left[\frac{2X_i}{\theta} - 2\theta \right] \\
&= -2n\theta + \sum_{i=1}^{n} \frac{2X_i}{\theta}
\end{aligned}
$$

således at maximum likelihood **estimatoren**, $\hat{\theta}_X$, kan findes som løsningen til ligningen

$$
\begin{gathered}
-2n\hat{\theta}_X + \sum_{i=1}^{n} \frac{2X_i}{\hat{\theta}_X} = 0 \\
\Updownarrow \\
n\hat{\theta}_X^2 = \sum_{i=1}^{n} X_i \\
\Updownarrow \\
\hat{\theta}_X = \sqrt{\frac{1}{n} \sum_{i=1}^{n} X_i}
\end{gathered}
$$

(hvor vi bruger, at parameter-rummet for θ er $\Theta = \{\mathbb{R} : \theta > 0\}$, hvorfor vi kan ignorere

løsningen $-\sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i}$).
Ved at bruge $n = 251$ og $\frac{1}{251}\sum_{i=1}^{251} x_i = 2.291$ kan vi udlede **estimatet** for vores data som

$$\begin{aligned}
\hat{\theta}_x = \hat{\theta}(x_1, \ldots, x_{251}) &= \sqrt{\frac{1}{251}\sum_{i=1}^{251} x_i} \\
&= \sqrt{2.291} \\
&\approx 1.514
\end{aligned}$$

3. Hesse-matricen er en skalar i dette tilfælde og givet ved den anden-afledte. Vi får at bidraget for borger i er

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

og dermed er informationen

$$\begin{aligned}
I(\theta_0) &= \mathbb{E}(-H_i(\theta_0)) \\
&= \mathbb{E}\left(\frac{2X_i}{\theta_0^2} + 2\right) \\
&= 2\frac{\mathbb{E}(X_i)}{\theta_0^2} + 2 \\
&= 2\frac{\theta_0^2}{\theta_0^2} + 2 \\
&= 4
\end{aligned}$$

Variansen bliver således

$$\begin{aligned}
Var(\hat{\theta}_X) &= \frac{1}{n} I(\theta_0)^{-1} \\
&= \frac{0.25}{n} \\
&\approx 0.00099
\end{aligned}$$

Det ses at standardafvigelsen bliver $se(\hat{\theta}_X) = \sqrt{Var(\hat{\theta}_X)} = \sqrt{0.00099} \approx 0.0316$.

4. Vi kan bruge den estimerede model til at beregne sandsynligheden for at en borger

sender mindst én ansøgning

$$P(X_i > 0) = 1 - P(X_i \leq 0) = 1 - \frac{\hat{\theta}^{2\cdot 0}}{0!}\exp(-\hat{\theta}^2) = 1 - \exp(-1.514^2) \approx 0.8989$$

5. Log-likelihood funktionen for den betingede model er

$$\begin{aligned}\log L_n(\theta,\delta) &= \log L(\theta,\delta|x_1,\dots,x_{251},d_1,\dots,d_{251})\\ &= \log\left(\prod_{i=1}^{251}\frac{(\theta+\delta d_i)^{2x_i}}{x_i!}\exp(-(\theta+\delta d_i)^2)\right)\\ &= \sum_{i=1}^{251}\{2x_i\log(\theta+\delta d_i)-(\theta+\delta d_i)^2-\log(x_i!)\}\end{aligned}$$

6. Hvis $\delta \neq 0$ så er der en tendens til at forsøget påvirker antallet af ansøgninger. Hvis $\delta > 0$ så er der en tendens til at deltagelse i forsøget får borgerne til at sende flere ansøgninger.

7. Vi kunne estimere den betingede model i STATA ved at skrive
```
mlexp ( 2*x*log({theta}+{delta}*d) -({theta}+{delta}*d)^2 )
```

8. Vi skal teste om der er en signifikant effekt af forsøget. Det svarer til hypotesen

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

med alternativ-hypotesen

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

Vi beregner vores z-statistik som

$$z_n(\delta = 0) = \frac{\hat{\delta} - 0}{se(\hat{\delta})} = \frac{0.1411508}{0.063545} \approx 2.2212731.$$

Vi ved at $z_n(\delta = 0) \overset{a}{\sim} \mathcal{N}(0,1)$ under $\mathcal{H}_0$. Så vi kan beregne den kritiske 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 afvise på et 5% signifikansniveau, at der IKKE er en effekt af forsøget. (p-værdien er $2 \cdot (1 - \Phi(2.221)) \approx 0.026$, hvilket er lavere end de 5%)
Alternativt kan LR-test benyttes da vi har fået log-likelihood funktionen under en restrikterede og urestrikterede model. Her får vi LR-statistikken

$$LR(\delta = 0) = 2(-95.916305 + 98.372667) = 4.9127$$

og vi ved at under $\mathcal{H}_0$, så er $LR \sim \chi_1^2$ med 1 frihedsgrad. Den kritiske værdi er således $F^{-1}_{\chi_1^2}(0.95) = 3.84$. Da $LR(\delta = 0) > 3.84$ kan vi her også afvise at der IKKE er en effekt af forsøget. p-værdien bliver næsten den samme $1 - F_{\chi_1^2}(4.9127) \approx 0.0263$.