Re-eksamen på Økonomistudiet vinter 2015-16
Sandsynlighedsteori og Statistik
2. Årsprøve
18. februar, 2016

(3-timers prøve med hjælpemidler)

Rettevejledning

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

Opgave 1

En virksomhed producerer træskåle. Produktionen består af to arbejdsprocesser, som udføres efter hinanden. Den tid, som bruges på den første proces, kan beskrives som en stokastisk variabel X_1 og tiden, som bruges på den anden proces, kan beskrives som X_2. Tabel 1 angiver fordelingen af tid.

Tabel 1: *Den simultane fordeling af tidsforbrug*

Arbejdstid i timer		Proces 1		
		1	2	3
	1	0.10	0.10	0.05
Proces 2	2	0.15	0.10	0.15
	3	0.20	0.10	0.05

1. Den marginale fordeling for tidsforbruget ved proces 1 kan beregnes ved at anvende sætning 4.2.1 (se Sørensen side 124)

$$\begin{aligned} P(X_1 &= 1) = p_1(1) = p(1,1) + p(1,2) + p(1,3) = 0.1 + 0.15 + 0.20 = 0.45 \\ P(X_1 &= 2) = p_1(2) = p(2,1) + p(2,2) + p(2,3) = 0.1 + 0.10 + 0.1 = 0.3 \\ P(X_1 &= 3) = p_1(3) = p(3,1) + p(3,2) + p(3,3) = 0.05 + 0.15 + 0.05 = 0.25 \end{aligned}$$

2. Der er flere måde at vise, at X_1 og X_2 ikke er uafhængige. En måde er at benytte definitionen af uafhængighed og vise at

$$P(X_1 = 1, X_2 = 1) \neq P(X_1 = 1) \cdot P(X_2 = 1).$$

Man kan vise, at $P(X_2 = 1) = 0.25$ og da $0.25 \cdot 0.45 = 0.112\,5 \neq 0.10$, kan X_1 og X_2 ikke være uafhængige (se sætning 3.6.1, Søresen side 82). Alternative måder at vise afhængighed er ved at vise, at $Cov(X_1, X_2) \neq 0$ eller $P(X_1 = 1|X_2 = 1) \neq P(X_1 = 1)$.

3. Kovariansen mellem de to tidsforbrug: $Cov(X_1, X_2)$ udregnes ved at benytte formel (3.8.2) Sørensen side 96. Derfor udregnes først

$$\begin{aligned} E(X_1X_2) &= 3.7 \\ E(X_1) &= 1.8 \\ E(X_2) &= 2.1 \end{aligned}$$

Det følger så at

$$\begin{aligned} Cov(X_1, X_2) &= E(X_1X_2) - E(X_1)E(X_2) \\ &= 3.7 - 1.8 \cdot 2.1 = -0.08 \end{aligned}$$

Kovariansen er altså negativ hvilket vil sige, at der er en negativ samvariation mellem de to arbejdsprocesser. Det tyder atså på at hvis den første del af processen er udført omhyggeligt (der er brugt lang tid på den) kan man spare lidt tid på anden del af processen.

4. Den samlede arbejdstid er givet ved $Z = X_1 + X_2$. Vi kan nu udregne sandsynligheden for at den samlede arbejdstid er 3 timer ved at benytte sætning 4.2.4 (se Sørensen side 126)

$$\begin{aligned} P(Z &= 3) \\ &= \sum_{j=0}^{3} p(j, y-j) \\ &= p(1,2) + p(2,1) = 0.15 + 0.10 = 0.25 \end{aligned}$$

Opgave 2

Kvaliteten af varen varierer og kan beskrives med et indeks mellem -1 og 5. Vi antager at indekset for kvalitet Q kan beskrives som en ligefordeling på intervallet -1 til 5. Varer, som har en kvalitet under 0, kasseres.

1. Da Q er en ligefordelt stokastisk variabel på intervallet -1 til 5 kan sandsynlighedstætheden for Q skrives som (se Eksempel 5.1.3, Sørensen side 148)

$$p(q) = \frac{1_{[-1,5]}(q)}{5 - -1} = \frac{1}{6} 1_{[-1,5]}(q).$$

   Sandsynligheden for at en vare må kaseres $P(Q \leq 0)$ kan bestemmes (jvf. Sørensen side 148) som

$$\begin{aligned} P(Q \quad \leq \quad 0) &= \int_{-\infty}^{0} \frac{1}{6} 1_{[-1,5]}(q)dq \\ &= \frac{1}{6} \int_{-1}^{0} 1_{[-1,5]}(q)dq = \frac{1}{6}[q]_{-1}^{0} = \frac{1}{6}[0 - -1] = \frac{1}{6} \end{aligned}$$

2. Den forventede kvalitet:

   (a) $E(Q)$ kan beregnes som middelværdien (se definition 5.2.1, Sørensen side 156, det kan bemærkes, at middelværdien er veldefineret, fordi Q er en begrænset stokastisk variabel):

$$\begin{aligned} E(Q) &= \int_{-\infty}^{\infty} q \frac{1}{6} 1_{[-1,5]}(q)dq \\ &= \int_{-1}^{5} q \frac{1}{6} dq \\ &= \frac{1}{6}[\frac{1}{2}q^2]_{-1}^{5} = \frac{1}{12}[25 - 1] = 2. \end{aligned}$$

   (b) Dernæst bestemmes den forventede kvalitet af de varer, som ikke kasseres: $E(Q|Q > 0)$. Denne betingede middelværdi kan udregnes ved at benytte formel C.3 i Rahbek side 5)

$$\begin{aligned} E(Q|Q \quad > \quad 0) = \frac{\int_0^5 q \frac{1}{6} 1_{[-1,5]}(q)dq}{P(Q > 0)} \\ &= \frac{\int_0^5 q \frac{1}{6}(q)dq}{1 - P(Q \leq 0)} = \frac{1}{\frac{5}{6}} \frac{1}{6}[\frac{1}{2}q^2]_0^5 = \frac{1}{5}\frac{1}{2}[25 - 0] = 2.5 \end{aligned}$$

3. Virksomheden kan sælge produktet til en pris, som er afhænger af den forventede kvalitet af produkterne som sælges: $p = 1000 \cdot *E(Q|Q > 0)$. Den forventede pris er $p = 1000 \cdot 2.5 = 2500$ kr. Indtjeningen for en varer som produceres er givet ved $p \cdot 1_{\{Q>0\}} + 0 \cdot 1_{\{Q \leq 0\}} = p \cdot 1_{\{Q>0\}}$. Den forventede indtjening:

$$\begin{aligned} E(p \cdot 1_{\{Q>0\}}) &= E(2500 \cdot 1_{\{Q>0\}}) = 2500 \cdot E(1_{\{Q>0\}}) \\ &= 2500 \cdot P(Q > 0) \\ &= 2500 \cdot \frac{5}{6} = 2083.3 \end{aligned}$$

4. Virksomheden overvejer nu at introducere en ny teknologi. Med den nye teknologi vil kvaliteten kunne beskrives som en ligefordeling på intervallet mellem -0.5 til 4. Ved tilsvarende udregninger som i spørgsmål 3 kan man udregne at

$$\begin{aligned} E(p \cdot 1_{\{Q_{NY}>0\}}) &= 1000 \cdot E(Q_{NY}|Q_{NY} > 0) \cdot P(Q_{NY} > 0) \\ &= 1000 \cdot 2 \cdot \frac{8}{9} = 1777.8 \end{aligned}$$

   Så selvom man kasserer færre varer ved den nye produktion vil indtjeningen ikke stige da man samtidigt få en forventet lavere kvalitet af de varer som ikke kasseres.

Opgave 3

Den gode opgave genkender tætheden som en exponential-fordeling, som er anvendt i kurset. Det vil muligvis lette udregningerne, men er ikke vigtigt for besvarelsen.

1. For at vise resultatet for fordelingsfunktionen, kan der fx differentieres,

$$\begin{aligned}\frac{\partial}{\partial y}F_{Y_i|X_i}(y \mid x;\theta) &= \frac{\partial}{\partial y}\left(1-\exp\left(\frac{-y}{x\cdot\theta}\right)\right)\\ &= (\theta\cdot x)^{-1}\exp\left(\frac{-y}{x\cdot\theta}\right)\\ &= f_{Y_i|X_i}(y \mid x;\theta).\end{aligned}$$

   Vi bemærker, at $F_{Y_i|X_i}(0 \mid x;\theta) = 0$ og $F_{Y_i|X_i}(y \mid x;\theta) \to 1$ når $y \to \infty$. Opgaven kan også løses ved at integrere tætheden.

2. Udgangspunktet med de betingede fordelinger bygger på en antagelse om *eksogenitet*.

   På grund af antagelsen om *identiske betingede fordelinger*, fås

$$\begin{aligned}\ell(\theta \mid y_i, x_i) &= f_{Y_i|X_i}(y_i \mid x_i;\theta)\\ &= (\theta\cdot x_i)^{-1}\exp\left(\frac{-y_i}{x_i\cdot\theta}\right).\end{aligned}$$

   Antagelsen om *uafhængige fordelinger* giver derfor

$$\begin{aligned}L(\theta \mid y_1,...,y_n,x_1,...,x_n) &= \prod_{i=1}^{n}\ell(\theta \mid y_i,x_i)\\ &= \prod_{i=1}^{n}(\theta\cdot x_i)^{-1}\exp\left(\frac{-y_i}{x_i\cdot\theta}\right).\end{aligned}$$

   Tilsvarende er log-likelihood bidraget givet ved

$$\begin{aligned}\log\ell(\theta \mid y_i,x_i) &= -\log(\theta\cdot x_i) - \frac{y_i}{x_i\cdot\theta}\\ &= -\log(\theta) - \log(x_i) - \frac{1}{\theta}\frac{y_i}{x_i}\end{aligned}$$

   og log-likelihood funktionen er

$$\begin{aligned}\log L(\theta \mid y_1,...,y_n,x_1,...,x_n) &= \sum_{i=1}^{n}\log\ell(\theta \mid y_i,x_i)\\ &= -n\log(\theta) - \sum_{i=1}^{n}\log(x_i) - \frac{1}{\theta}\sum_{i=1}^{n}\frac{y_i}{x_i}.\end{aligned}$$

3. For at udlede estimatet som funktion af observationerne, findes først score-bidraget,

$$\begin{aligned}s_i(\theta) &= \frac{\partial}{\partial\theta}\log\ell(\theta \mid y_i,x_i)\\ &= \frac{\partial}{\partial\theta}\left(-\log(\theta)-\log(x_i)-\frac{1}{\theta}\frac{y_i}{x_i}\right)\\ &= -\frac{1}{\theta}+\frac{1}{\theta^2}\frac{y_i}{x_i},\end{aligned}$$

   sådan at scoren er givet ved

$$\begin{aligned}S(\theta) &= \sum_{i=1}^{n}s_i(\theta)\\ &= \frac{1}{\theta^2}\sum_{i=1}^{n}\frac{y_i}{x_i}-\frac{n}{\theta}.\end{aligned}$$

Første-ordens betingelsen giver derfor et estimat,

$$S(\hat{\theta}_n) = 0$$

$$\frac{1}{\hat{\theta}_n^2} \sum_{i=1}^{n} \frac{y_i}{x_i} - \frac{n}{\hat{\theta}_n} = 0,$$

sådan at

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{x_i}.$$

4. Fra skewness og kurtosis i tabellen med beskrivende statistik, lader det til at x kunne være normalfordelt.

   I vores tilfælde er estimatet givet ved, $\hat{\theta}_n = 5.275$.

5. Dette spørgsmål er relativt teknisk og man kan næppe vente helt samme præcision som i besvarelsen her.

   (a) Bidraget til Hessematricen fra observation i er givet ved

   $$\begin{aligned} H_i(\theta) &= \frac{\partial}{\partial \theta} s_i(\theta) \\ &= \frac{\partial}{\partial \theta} \left(-\frac{1}{\theta} + \frac{1}{\theta^2} \frac{Y_i}{X_i} \right) \\ &= \frac{1}{\theta^2} - \frac{2}{\theta^3} \frac{Y_i}{X_i}. \end{aligned}$$

   (b) Informationen findes defor ved direkte udregning,

   $$\begin{aligned} I(\theta) &= -E\left[H_i(\theta) \mid X_i = x_i\right] \\ &= -E\left[\frac{1}{\theta^2} - \frac{2Y_i}{\theta^3 X_i} \mid X_i = x_i\right] \\ &= -\frac{1}{\theta^2} + \frac{2}{\theta^3} E\left[\frac{Y_i}{X_i} \mid X_i = x_i\right] \\ &= -\frac{1}{\theta^2} + \frac{2}{\theta^3} \frac{E\left[Y_i \mid X_i = x_i\right]}{x_i}. \end{aligned}$$

   Vi bruger nu at $E(Y_i \mid X_i = x_i) = x_i \cdot \theta$, sådan at

   $$\begin{aligned} I(\theta) &= \frac{2}{\theta^3} \frac{x_i \cdot \theta}{x_i} - \frac{1}{\theta^2} \\ &= \frac{1}{\theta^2}. \end{aligned}$$

   (c) Variansen på estimatoren bliver derfor

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

   og $\mathsf{se}(\hat{\theta}_n) = \theta_0/\sqrt{n}$.

   (d) I vores tilfælde fås, med indsatte estimatorer,

   $$V(\hat{\theta}_n) = \frac{\theta_0^2}{n} \approx \frac{\hat{\theta}_n^2}{n} = \frac{5.275^2}{200} = 0.139$$

   og

   $$\mathsf{se}(\hat{\theta}_n) = \sqrt{V(\hat{\theta}_n)} = \sqrt{0.139} = 0.373.$$

6. Her skal bruges fordelingsfunktionen fra spørgsmål 1. Vi ved, at

$$F_{Y_i|X_i}(y \mid x; \theta) = 1 - \exp\left(\frac{-y}{x \cdot \theta}\right).$$

Sandsynligheden kan derfor udregnes som

$$\begin{aligned} P(Y > 50) &= 1 - P(Y \leq 50) \\ &= 1 - F_{Y_i|X_i}(50 \mid x = 5; \hat{\theta}_n) \\ &= 1 - (1 - \exp\left(\frac{-50}{5 \cdot \theta}\right)) \\ &= \exp\left(\frac{-50}{5 \cdot 5.275}\right) \\ &= 0.150, \end{aligned}$$

svarende til 15 procent.

7. Hypotesen om $\theta_0 = 5$ opstilles som

$$H_0 : \theta_0 = 5 \quad \text{mod} \quad H_A : \theta_0 \neq 5.$$

Det kan testes med et z-teststørrelse,

$$z_n(\theta_0 = 5) = \frac{\hat{\theta}_n - 5}{\mathsf{se}(\hat{\theta}_n)} = \frac{5.275 - 5}{0.373} = 0.737.$$

Teststørrelsen skal sammenlignes med en kritisk værdi fra en stardard normalfordeling. På et 5% signifikans-niveau er den kritiske værdi $\Phi^{-1}(0.975) = 1.96$ og hypotesen om $\theta_0 = 5$ kan ikke forkastes.