E6 opgavesæt 3

Jogvan M. Poulsen

30. november 2000

Opgave 1 (sommer 94, opgave 3)

Gonorré-tilfælde Danmark 1989-91
opdelt efter smitteår og de smittedes alder og køn

Køn	Alder	1989	1990	1991	Ialt
	16-19	105	88	46	239
		(110,02)	(79,58)	(49,40)	
Mænd	20-24	447	297	195	939
		(432,24)	(312,66)	(194,10)	
	25-	466	351	220	1037
		(477,35)	(345,29)	(214,36)	
	Ialt	1018	736	461	2215
	16-19	205	132	95	432
		(198,86)	(143,84)	(89,30)	
Kvinder	20-24	288	237	124	649
		(298,75)	(216,10)	(134,15)	
	25-	177	116	78	371
		(170,78)	(123,53)	(76,69)	
	Ialt	670	485	297	1452
	Total	1688	1221	758	3667

Opgave a Vi skal vise, at den tosidede tabel man får frem ved at betragte kombinationen af køn og alder på den ene side og år på den anden side kan beskrives ved en multiplikativ Poissonmodel. Vi skal med andre ord vise at der eksisterer $(\psi, \alpha, \beta) \in [0, \infty[\times \Delta_k \times \Delta_m$, så at $\lambda_{rs} = \psi \alpha_r \beta_s$.

Statistisk model. Vi antager at variablen X_{rs} er Poissonfordelt med parameter λ_{rs}, hvor variablene X_{rs} kan samles i en tosides tabel

$$
\begin{array}{|ccc|c|}
\hline
X_{11} & \cdots & X_{1m} & X_{1\cdot} \\
\vdots & & \vdots & \vdots \\
X_{k1} & \cdots & X_{km} & X_{k\cdot} \\
\hline
X_{\cdot 1} & \cdots & X_{\cdot m} & X_{\cdot\cdot} \\
\hline
\end{array}
$$

Det antages yderligere, at variablene X_{rs} er stokastisk uafhængige, (selv om man nok kunne forvente det modsatte, altså at der er flere mænd der smitter kvinder end kvinder der smitter kvinder) hvor i vores konkrete tilfælde $m = 3$ og $k = 6$ Med et udfaldsrum $\mathbf{N}_0^{km}$ bliver den statistiske model bliver

$$(\mathbf{N}_0^{18}, (P_{(\lambda_{11} \cdots \lambda_{18})})_{(\lambda_{11} \cdots \lambda_{18}) \in [0,\infty[^{18}})$$

hvor fordelingen er givet ved

$$P_{(\lambda_{11} \cdots \lambda_{18})}(X = x) = \prod_{r=1}^{6} \prod_{s=1}^{3} \frac{\lambda_{rs}^{x_{rs}}}{x_{rs}!} e^{-\lambda_{rs}}$$

Maksimaliseringsestimator ML $(\hat{\lambda}_{11}, \cdots, \hat{\lambda}_{18}) = (\hat{\lambda}_{11}, \cdots, \hat{\lambda}_{18}(x))$ er givet ved $\hat{\lambda}_{rs} = x_{rs}$ for $r = 1, \cdots, 18$ og $s = 1, \cdots, 3$, hvor fordelingen er

$$P_{(\lambda_{11} \cdots \lambda_{18})}(\hat{\lambda}_{11} = x_{11}, \cdots, \hat{\lambda}_{18} = x_{18}) = \prod_{r=1}^{6} \prod_{s=1}^{3} \frac{\lambda_{rs}^{x_{rs}}}{x_{rs}!} e^{-\lambda_{rs}}$$

Vi skal nu opstille en nulhypotese.

$$\begin{aligned} H_0 \quad : \quad & \lambda_{rs} = \psi \alpha_r \beta_s \qquad r = 1, \cdots, 6 \quad , s = 1, \cdots, 3 \\ & (\psi \alpha \beta) \in [0, \infty[\times \Delta_6 \times \Delta_3 \end{aligned}$$

ML under nullhypotesen er givet ved

$$\begin{aligned} \hat{\psi} &= x_{\cdot\cdot} = 3667 \\ \hat{\alpha}_r &= \frac{x_{r\cdot}}{x_{\cdot\cdot}} \approx (0,065; 0,256; 0,283; 0,118; 0,177; 0,101) \\ \hat{\beta}_s &= \frac{x_{\cdot s}}{x_{\cdot\cdot}} \approx (0,460; 0,333; 0,208) \end{aligned}$$

Vi kan nu udregne værdierne $\hat{\psi}\hat{\alpha}_r\hat{\beta}_s = {}^{x_{r\cdot} x_{\cdot s}}/_{x_{\cdot\cdot}}$. De er indsat i tabellen i parantes. Nu kommer det afgørende øjeblik, hvor vi skal beregne testsandsynligheden. Vi har at

$$-2 \ln Q(x) = -2 \sum_{r=1}^{2} \sum_{m=1}^{3} x_{rs} \ln \frac{x_{r\cdot} x_{\cdot s}}{x_{rs} x_{\cdot\cdot}} \approx 8,548$$

Dette er χ^2-fordelt med $f = (6-1)(3-1) = 10$ frihedsgrader. Ved at se i tabellen i TT ser vi, at siden $8,548 < 18,307$ kan vi konkludere, at vi ikke kan afvise at vores hypotese er rigtig.

Opgave b Her skal vi svare på spørgsmålet, om en multiplikativ model kan beskrive antal gonorr,-tilfældes afhængighed af alder og køn.

Vi laver samme procedure som i opgave a, hvor vi samler alle tilfældene pr år. Dette kan vi tillade os, fordi vi lige har vist, at gonorr,tilfældene, ikke afhang af årstallet. Den statistiske model er den samme som før, vi har bare at $k = 3$ og $m = 2$.

Gonorr,-tilfælde Danmark 1989-91
opdelt efter de smittedes alder og køn

Alder	Mand	Kvinde	Ialt
16-19	239	432	671
	(405,31)	(265,69)	
20-24	939	649	1588
	(959,21)	(628,79)	
25-	1037	371	1408
	(850,48)	(557,52)	
Ialt	2215	1452	3667

Vi skal nu opstille en tilsvarende nulhypotese.

$$\begin{aligned} H_0 \quad : \quad & \lambda_{rs} = \psi \alpha_r \beta_s \qquad r = 1, \cdots, 3 \quad , s = 1, \cdots, 2 \\ & (\psi \alpha \beta) \in [0, \infty[\times \Delta_3 \times \Delta_2 \end{aligned}$$

ML under nullhypotesen er givet ved

$$\begin{aligned}
\hat{\psi} &= x_{\cdot\cdot} = 3667 \\
\hat{\alpha}_r &= \frac{x_{r\cdot}}{x_{\cdot\cdot}} \approx (0,183; 0,433; 0,384) \\
\hat{\beta}_s &= \frac{x_{\cdot s}}{x_{\cdot\cdot}} \approx (0,604; 0,396)
\end{aligned}$$

Vi kan nu udregne værdierne $\hat{\psi}\hat{\alpha}_r\hat{\beta}_s = {}^{x_{r\cdot}x_{\cdot s}}/_{x_{\cdot\cdot}}$. De er indsat i tabellen i parantes. Nu beregner vi testsandsynligheden.

$$-2\ln Q(x) = -2\sum_{r=1}^{2}\sum_{m=1}^{3} x_{rs}\ln\frac{x_{r\cdot}x_{\cdot s}}{x_{rs}x_{\cdot\cdot}} \approx 277,61$$

Dette er χ^2-fordelt med $f = (3-1)(2-1) = 2$ frihedsgrader. Ved at se i tabellen i TT ser vi, at siden $277,61 >> 18,421$ kan vi konkludere, at vores hypotese overhoved ikke holder. Man kan således ikke ikke beskrive antal gonorr,-tilfældes afhængighed af alder og kom med en multiplikativ model.

Opgave 2 (sommer 94, opgave 2)

Opgave a Vi skal vise, af funktionen givet ved

$$f(x) = \begin{cases} \frac{1}{\sqrt{2\pi}}\frac{1}{x}e^{-\frac{1}{2}(\ln(x))^2} & \text{for } x > 0 \\ o & \text{ellers} \end{cases}$$

er en sandsynlighedstæthed, og at $Y = \ln X$ er normeret normalfordelt. For at $f(x)$ skal være en sandsynlighedstæthed, skal den opfylde to betingelser, nemlig

$$\begin{aligned}
f(x) &\geq 0 \\
\int_E f(x)dx &= 1
\end{aligned}$$

hvor $E = [0, \infty[$ er udfaldsrummet. Det det ses let, at for alle $x \geq 0$ bliver $f(x) \geq 0$. For at vise den anden betingelse, har jeg anvent 'Derive', og er kommet frem til at

$$\int_0^\infty \frac{1}{\sqrt{2\pi}}\frac{1}{x}e^{-\frac{1}{2}(\ln(x))^2}dx = 1$$

Det betyder, at $f(x)$ er en sandsynlighedstæthed.

For at vise, at $Y = \ln X$ er normeret normalfordelt, skal vi bruge afbildningen $t(x) = \ln x$ og $t'(x) = {}^1/_x$, samt at $t^{-1}(y) = e^y$. Herved får vi

$$\begin{aligned}
q(y) &= \frac{f(t^{-1}(y))}{|t'(t^{-1}(y))|} \\
&= \frac{f(e^y}{|t'(e^y)|} \\
&= \frac{\frac{1}{\sqrt{2\pi}}\frac{1}{e^y}e^{-\frac{1}{2}(\ln(e^y))^2}}{\frac{1}{|e^y|}} \\
&= \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2}
\end{aligned}$$

Denne funktion med middelværdi 0 og varians 1, er den normerede normalfordeleling (Gaussfordeling).

Opgave b Hvis X_1 og X_2 er uafhængige med samme fordeling med den før givne tæthed, og $\alpha \neq \beta$ er to reelle tal, samt at de afledte stokastiske variable W og Z givet ved

$$W = X_1^\alpha X_2^\beta \qquad \text{og} \qquad Z = X_1 X_2$$

Da skal vi vise, ved at transformere med logaritmen, at W og Z er uafhængige hvis og kun hvis $\alpha + \beta = 0$. Hvis $\ln W$ og $\ln Z$ er uafhængige, da er W og Z det også, hvilket betyder, at covariansen mellem dem er nul. Jeg mener ikke det glæder den anden vej, men jeg skal ikke vise at W og Z er uafhængige altid, men når de er det, er det kun når covariansen er nul, altså $\alpha + \beta = 0$.

$$\begin{aligned}
\ln Z &= \ln(X_1 X_2) = \ln X_1 + \ln X_2 \\
\ln W &= \ln(X_1^\alpha X_2^\beta) = \alpha \ln X_1 + \beta \ln X_2
\end{aligned}$$

Covariansen er givet ved:

$$\begin{aligned}
\operatorname{cov}(\ln Z, \ln W) &= \mathrm{E}(\ln Z \cdot \ln W) - (\mathrm{E} \ln Z) \cdot (\mathrm{E} \ln W) \\
&= \mathrm{E}((\ln X_1 + \ln X_2)(\alpha \ln X_1 + \beta \ln X_2)) \\
&\quad -(\mathrm{E}(\ln X_1 + \ln X_2))(\mathrm{E}(\alpha \ln X_1 + \beta \ln X_2)) \\
&= \alpha \mathrm{E}(\ln X_1)^2 + \beta \mathrm{E}(\ln X_2)^2 + (\alpha + \beta)\mathrm{E}(\ln X_1 \ln X_2) \\
&\quad -(\mathrm{E}(\ln X_1) + \mathrm{E}(\ln X_2))(\alpha \mathrm{E}(\ln X_1) + \beta \mathrm{E}(\ln X_2)) \\
&= \alpha \mathrm{E}(\ln X_1)^2 + \beta \mathrm{E}(\ln X_2)^2 + (\alpha + \beta)\mathrm{E}(\ln X_1 \ln X_2) \\
&\quad -(\alpha(\mathrm{E} \ln X_1)^2 + \beta(\mathrm{E} \ln X_2)^2 + (\alpha + \beta)(\mathrm{E}(\ln X_1)\mathrm{E}(\ln X_2)) \\
&= \alpha(\mathrm{E}(\ln X_1)^2 - (\mathrm{E} \ln X_1)^2) + \beta(\mathrm{E}(\ln X_2)^2 - (\mathrm{E} \ln X_2)^2) \\
&\quad +(\alpha + \beta)(\mathrm{E}(\ln X_1 \ln X_2) - \mathrm{E}(\ln X_1)\mathrm{E}(\ln X_2))
\end{aligned}$$

Siden X_1 og X_2 har samme fordeling, kan vi erstatte X_2 med X_1, og vi får

$$\begin{aligned}
&= \alpha(\mathrm{E}(\ln X_1)^2 - (\mathrm{E} \ln X_1)^2) + \beta(\mathrm{E}(\ln X_1)^2 - (\mathrm{E} \ln X_1)^2) \\
&\quad +(\alpha + \beta)(\mathrm{E}(\ln X_1)^2 - (\mathrm{E} \ln X_1)^2) \\
&= 2(\alpha + \beta)(\mathrm{E}(\ln X_1)^2 - (\mathrm{E} \ln X_1)^2) \\
&= 2(\alpha + \beta)\mathrm{Var}(X_1)
\end{aligned}$$

Variansen af X_1 bliver aldrig nul, derfor er covariansen kun nul når $\alpha + \beta = 0$. Vi har nu vist, at når $\alpha + \beta = 0$ er det muligt at W og Z er stokastisk uafhængige, men ikke en selvfølge. Er derimod $\alpha + \beta \neq 0$ er de med garanti ikke stokastisk uafhængige.

Opgave 3

Vi får givet to positive reelle stokastiske variable X og Y med simultan fordeling givet ved en tæthed $p(x, y)$, hvor $x, y > 0$. De to afledte stokastiske variable S og Z er defineret til

$$S = X + Y \qquad \text{og} \qquad Z = \frac{X}{X + Y}$$

Opgave a Vi skal finde tætheden $q(s,z)$ for (S,Z)

Siden vi skal finde tætheden $q(s,z)$ udfra variablene S og Z, og tætheden $p(x,y)$, definerer først transformationen $t:\mathbf{R}^2\to\mathbf{R}^2$ på følgende måde

$$\begin{aligned} t(x,y) &= (x+y, \frac{x}{x+y}) = (s,z) \\ t^{-1}(s,z) &= (sz, s-sz) = (x,y) \end{aligned}$$

Funktionsmatricen Dt skal vi også bruge, og den er givet ved

$$Dt = \begin{pmatrix} \frac{\partial(x+y)}{\partial x} & \frac{\partial(x+y)}{\partial y} \\ \frac{\partial(\frac{x}{x+y})}{\partial x} & \frac{\partial(\frac{x}{x+y})}{\partial y} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ \frac{y}{(x+y)^2} & \frac{-x}{(x+y)^2} \end{pmatrix}$$

Determinanten følger heraf

$$\left|\det\begin{pmatrix} 1 & 1 \\ \frac{y}{(x+y)^2} & \frac{-x}{(x+y)^2} \end{pmatrix}\right| = \left|\frac{-x}{(x+y)^2} - \frac{y}{(x+y)^2}\right| = \frac{|-x-y|}{(x+y)^2}$$

Vi kan se, at $x=y \Rightarrow \det(Dt)=0$ som betyder, at vi ikke har en bijektiv afbildning fra $x,y\to s,z$, medmindre vi definerer at $x\neq y$. Før vi kan anvende sætning 6,2,2TT, skal vi udtrykke $\det(Dt)$ i s,z. Ved at bruge transformationen t^{-1} får vi, at

$$\frac{|-x-y|}{|(x+y)^2|} = \frac{|-sz-(s-sz)|}{|(sz+s-sz)^2|} = \frac{|-s|}{|s^2|} = \frac{1}{|s|}$$

Nu kan vi indsætte, hvor vi får

$$\begin{aligned} q(s,z) &= \frac{p(t^{-1}(s,z))}{|\det Dt(t^{-1}(s,z))|} \\ &= \frac{p(t^{-1}(s,z))}{\frac{1}{|s|}} \end{aligned}$$

Opgave b Nu skal vi finde tætheden for fordelingen af Z. Det gør vi ved at integrere tætheden for s ud (Siden X og Y altid var positive, integrerer vi over $\mathbf{R}_+$)

$$q(z) = \int_0^\infty p(t^{-1}(s,z))s\,ds$$

Her kan vi ikke udregne integralet, da vi ikke har noget konkret udtryk for tætheden $p(x,y)$.

Opgave c Fordelingen af Z, hvis X og Y er stokastisk uafhængige, normeret exponentialfordelte.

Det vil sige, at X's hhv. Y's tæthed kan opskrives som

$$p(x) = e^{-x} \qquad p(y) = e^{-y}$$

Når X og Y er stokastisk uafhøngige, er tætheden $p(x,y)=p(x)p(y)$.

$$p(x,y) = e^{-(x+y)}$$

Nu kan vi finde udtrykket $p(t^{-1}(s,z))$.

$$p(t^{-1}(s,z)) = e^{-s}$$

Vi indsætter nu i udtrykket fra spm. b

$$\begin{aligned} q(z) &= \int_0^\infty p(t^{-1}(s,z))sds \\ &= \int_0^\infty e^{-s}sds \\ &= \Gamma(2) = (1!) = 1 \end{aligned}$$

Det fortæller os on Z's fordeling, at den ligger i intervallet $[0,1]$

Opgave d Fordelingen af Z, hvis X og Y er stokastisk uafhængige, normeret exponentialfordelte med skalaparameter α henholdsvis β. Her kommer tæthedsfunktionen til at se ud som

$$p(x)p(y) = p(x,y) = \frac{1}{|\alpha\beta|}e^{-(\frac{x}{\alpha}+\frac{y}{\beta})}$$

Udtrykket $p(t^{-1}(s,z))$ kommer her til at se ud som

$$p(t^{-1}(s,z)) = \frac{1}{|\alpha\beta|}e^{-(\frac{sz}{\alpha}+\frac{s-sz}{\beta})}$$

For at finde fordelingen af Z Nu skal vi integrere s ud (spm. b)

$$\begin{aligned} q(z) &= \int_0^\infty p(t^{-1}(s,z))sds \\ &= \frac{1}{|\alpha\beta|}\int_0^\infty e^{-(\frac{sz}{\alpha}+\frac{s-sz}{\beta})}sds \\ &= \frac{1}{|\alpha\beta|}\int_0^\infty e^{-(s\frac{\alpha z+\alpha-\beta z}{\alpha\beta})}sds \end{aligned}$$

Her stopper nok festen for mit vedkommende må jeg erkende.