E6 opgavesæt 2

Jogvan M. Poulsen

6. november 2000

Opgave 2.28 i Statistik

I Københavns kommune blev der i året 1977 født 2560 drenge og 2401 piger. Samme år fødtes der i Stenløse kommune 91 drenge og 69 piger.Dette kan vi sammenfatte i et skema.

Fødsler	Dreng	Pige	ialt
København	2560	2401	4961
	x_K	$n_K - x_K$	n_K
Stenløse	81	69	150
	x_S	$n_S - x_S$	n_S

Hvor X_K angiver antallet af drengefødsler i københavn og X_K angiver antallet at antallet af drengefødsler i Stenløse. Vi antager at X_K og X_S er stokastisk uafhængige, dvs. ingen tvillingefødsler, ingen bliver talt med to gange osv.

Opgave a Her skal vi opstille en statistisk model, der kan beskrive forsøget og estimere sandsynligheden, p_K, for at føde en dreng i København.

Der må være tale om en bimomialfordeling, hvor vi ser på alle dem der har føde et barn, og hvor de enten har født en dreng, eller ikke en dreng (pige) Udfaldsrummet må være antal fødsler ialt. Dvs. $n_K = 4961$

$$E_K = \{0, 1, \cdots, 4961\}$$

Den statistiske model kan nu opstilles

$$(E_K, (P_{p_K})_{p_K \in [0,1]})$$

hvor punktsandsynligheden er givet ved

$$P_{p_K}(X_K = x_K) = \binom{n_K}{x_K} (p_K)^{x_K} (1 - p_K)^{n_K - k_K}$$

Det eneste vi ved om fødslerne i København 1977, er de tal der står i tabellen. Derfor må vi formode, at sandsynligheden p_K ligger tæt på maksimaliseringsestimatoren $\hat{p}_K$, som for en binomialfordeling er entydig bestemt og givet ved.

$$\hat{p}_K = \frac{x_K}{n_K} = \frac{2560}{4961} \approx 0,516$$

Vi kan så opskrive sandsynlighedsfunktionen med $p_K = \hat{p}_K$ for $x_K \in E_K$

$$P(X_K = x_K) = \binom{4961}{x_K} (0,516)^{x_K} (0,484)^{4961 - x_K}$$

Opgave b Vi skal svare på, om man kan tillade sig at antage, at der fødes lige mange drenge som piger, altså $p_K = 0,5$?

Her har vi en såkaldt simpel nulhypotese hvor

$$H_0 : \hat{p}_{K0} = 0,5$$

Nu vil vi finde kvotienttesteststørrelsen

$$\begin{aligned} Q(x_K) &= \frac{L(x_K, \hat{p}_{K0}(x_K))}{L(x_K, \hat{p}_K(x_K))} \\ &= \frac{P_{\hat{p}_{K0}(x_K)}(X_K = x_K)}{P_{\hat{p}_K(x_K)}(X_K = x_K)} \\ &\approx \frac{\binom{4961}{2560}(0,5)^{2560}(0,5)^{4961-2560}}{\binom{4961}{2560}(0,516)^{2560}(1-0,516)^{4961-2560}} \\ &= \left(\frac{0,5}{0,516}\right)^{2560}\left(\frac{0,5}{0,484}\right)^{2401} \\ &\approx 0,078 \end{aligned}$$

Nu skal vi finde χ^2_1-fordelingen ved at $\chi^2_1 \approx -2\ln Q = 5,1$, hvor dimentionen for $\dim(p_k \in [1,0]) = 1$ og siden vi har en simpel nulhypotese er $\dim(\hat{p}_k) = 1$. Dette betyder at sandsynligheden ligger lige omkring under 5% hvilket betyder at vi, med et signfikantnevau på 5%, ikke kan konkludee at vores hypoteso om at sandsynligheden for at få en dreng i København i 1977 var 0,5 holder.

Fordelingen af $\hat{p}_K$ er givet ved

$$P_{\hat{p}_K}(\hat{p}_K = \frac{y}{n_K}) = \binom{n_K}{y}(\hat{p}_K)^y(1-\hat{p}_K)^{n_K-y}$$

hvor værdiområdet for $\hat{p}_K$ er

$$\hat{p}_K = \{0, \frac{1}{n_K}, \cdots, 1\}$$

Opgave c Nu skal vi forsøge at estimere sandsynligheden for drengefødsel på grundlag af tallene fra Stenløse

Vi anvender tilsvarende statistisk model som ved fødslerne i København.

$$E_S = \{0, 1, \cdots, 160\}$$

Den statistiske model for drengefødsler i Stenløse er derfor

$$(E_S, (P_{p_S})_{p_S \in [0,1]})$$

For at estimere sandsynligheden p_S finder vi maksimaliseringsestimatoren $\hat{p}_S$,

$$\hat{p}_S = \frac{x_S}{n_S} = \frac{81}{150} = 0,54$$

sandsynlighedsfunktionen med $p_S = \hat{p}_S$ for $x_S \in E_S$ bliver så

$$P(X_S = x_S) = \binom{160}{x_S}(0,54)^{x_S}(0,46)^{150-x_S}$$

Opgave d Hvad ville resultatet blive, hvis vi testede om sandsynligheden for drengefødsel var $0,5$ på grundlag af tallene fra Stenløse

Vi har igen en simpel nulhypotese

$$H_0 : \hat{p}_{S0} = 0,5$$

kvotienttesteststørrelsen bliver her

$$\begin{aligned} Q(x_S) &= \frac{L(x_S, \hat{p}_{S0}(x_S))}{L(x_S, \hat{p}_S(x_S))} \\ &= \frac{P_{\hat{p}_{S0}(x_S)}(X_S = x_S)}{P_{\hat{p}_S(x_S)}(X_S = x_S)} \\ &= \frac{\binom{150}{81}(0,5)^{81}(0,5)^{150-81}}{\binom{150}{81}(0,54)^{81}(1-0,54)^{150-81}} \\ &= \left(\frac{0,5}{0,54}\right)^{81}\left(\frac{0,5}{0,46}\right)^{69} \\ &\approx 0,62 \end{aligned}$$

Vi skal nu igen

Vi finder igen χ^2_1-fordelingen ved at $\chi^2_1 \approx -2\ln Q = 0,956$, hvor dimentionen igen er 1. Her ligger værdien langt under det vi kan slå op i TT, hvilket betyder, at sandsynligheden ligger godt 5% hvilket betyder at vi, med et signfikantnevau på 5%, ikke kan afvise vores hypoteso om at sandsynligheden for at få en dreng i Stenløse i 1977 var 0,5.

Fordelingen af $\hat{p}_S$ er givet ved

$$P_{\hat{p}_S}(\hat{p}_S = \frac{y}{n_S}) = \binom{n_S}{y}(\hat{p}_S)^y(1-\hat{p}_S)^{n_S-y}$$

hvor værdiområdet for $\hat{p}_S$ er

$$\hat{p}_S = \{0, \frac{1}{n_S}, \cdots, 1\}$$

Opgave 2.32 i Statistik

Nedestående data stammer fra en arbejdsmiljøundersøgelse på en keramisk virksomhed. Her undersøgtes bl.a. forekomsten af hudlidelser hos henholdsvis mænd og kvinder på to afdelinger.

med hudlidelse

Afdeling Køn	A	B
Mænd	5	6
Kvinder	10	18

uden hudlidelse

Afdeling Køn	A	B
Mænd	9	52
Kvinder	30	85

På grundlag af dette materiale, skal vi analysere og overveje hyppigheden af hudlidelser på de to afdelinger, og hvad der ville ske, hvis man så bort fra køn i analysen.

Jeg vælger at undersøge hyppigheden af hudlidelser kønnene hver for sig, og derefter se om de fordelinger jeg får med sandsynlighedsparameter hver for sig p_k, p_m ligger indenfor det tilladte med hensyn til den samlede fordeling med ssh p

Mænd	A	B	ialt
med	5	6	11
	x_a	x_b	$x.$
uden	9	52	61
	$n_a - x_a$	$n_b - x_b$	$n. - x.$
ialt	14	58	72
	n_a	n_b	$n.$

Hvor X_a angiver antal mænd med hudlidelser på andeling A, og X_b angiver antal mænd med hudlidelser på andeling B. Vi antager, at X_a og X_b er stokastisk uafhængige, dvs. ingen har arbejdet på bægge afdelinger, samt at ingen har valgt den ene afdeling frem for den anden af helbredsmæssige grunde.

Vi antager at X_a er binomialfordelt med (n_a, p_{ma}), og at X_b er binomialfordelt med (n_b, p_{mb}). Den statistiske model bliver derfor

$$(E_m, (P_{(p_{ma},p_{mb})})_{(p_{ma},p_{mb})\in[0,1]^2})$$

hvor udfaldsrummet er givet ved

$$E_m = \{0, 1, \cdots, 14\} \times \{0, 1, \cdots, 58\}$$

Fordelingen er dermed givet ved

$$P(X_a, X_b) = \binom{n_a}{x_a}(p_{ma})^{x_a}(1-p_{ma})^{n_a-x_a} \cdot \binom{n_b}{x_b}(p_{mb})^{x_b}(1-p_{mb})^{n_b-x_b}$$

Nu skal vi finde Likelihoodfunktionens største værdi for vores observation. Det gør vi ved at indsætte $\hat{p}_{ma} = {}^{x_a}/_{n_a}$ og tilsvarende for $\hat{p}_{mb}$. Vi får

$$(\hat{p}_{ma}(x_a), \hat{p}_{mb}(x_b)) = \left(\frac{x_a}{n_a}, \frac{x_b}{n_b}\right) = \left(\frac{5}{14}, \frac{6}{58}\right)$$

Vi antager nu, at $p_m = p_{ma} = p_{mb}$. Vi opstiller så en nulhypotese

$$H_0 : p_m = p_{ma} = p_{mb} \in [0, 1]$$

Under nulhypotesen bliver Likelihoodfunktionen

$$\begin{aligned} L(x_a, x_b, p_m) &= \binom{n_a}{x_a} p_m^{x_a}(1-p_m^{n_a-x_a}) \cdot \binom{n_b}{x_b} p_m^{x_b}(1-p_m^{n_b-x_b}) \\ &= \binom{n_a}{x_a}\binom{n_b}{x_b} p_m^{x.}(1-p_m)^{n.-x.} \end{aligned}$$

For at finde ML-estimatoren $\hat{P}_{m0}$, skal vi maksimere $L(x_a, x_b, p_m)$, Det kan vi med fordel se i eksempel 2.3.1IH, hvor der står at

$$\hat{p}_{m0}(x_a, x_b) = \frac{x.}{n.} = \frac{11}{72} \approx 0,153$$

Nu kommer så det store øjeblik, hvor vi skal se om vi kunne tillade os at antage at $p_{ma} = p_{mb}$. Dette gøres ved at finde kvotienttesteststørrelsen

$$Q(x_a, x_b) = \frac{L(x_a, x_b, \hat{p}_{m0})}{L(x_a, x_b, \hat{p}_{ma}, \hat{p}_{mb})}$$

$$
\begin{aligned}
&= \frac{\binom{n_a}{x_a}\binom{n_b}{x_b}(\hat{p}_m)^{x_\cdot}(1-\hat{p}_m)^{n_\cdot - x_\cdot}}{\binom{n_a}{x_a}(\hat{p}_{ma})^{x_a}(1-\hat{p}_{ma})^{n_a-x_a}\binom{n_b}{x_b}(\hat{p}_{mb})^{x_b}(1-\hat{p}_{mb})^{n_b-x_b}} \\
&\approx \frac{\left(\frac{5+6}{14+58}\right)^{x_\cdot}\left(1-\frac{5+6}{14+68}\right)^{n_\cdot - x_\cdot}}{\left(\frac{5}{14}\right)^{x_a}\left(1-\frac{5}{14}\right)^{n_a-x_a}\left(\frac{6}{58}\right)^{x_b}\left(1-\frac{6}{58}\right)^{n_b-x_b}} \\
&\approx \frac{(0,1528)^{11}(0,8472)^{61}}{(0,3571)^5(0,6429)^9(0,1034)^6(0,8966)^{52}} \\
&\approx 0,0940 \\
&\Rightarrow -2\ln Q \approx 4,729
\end{aligned}
$$

Med 1 frihedsgrad, giver χ^2-dordelingen os lige under 5%. Her kan vi foreløbig slutte, at vi ikke afvise, at sansynligheden for at få en hudlidelse for en mand på de respektive afdelinger er den samme.

Tilsvarende kan vi udregne kvotienttesstørrelsen for kvinderne på arbejdspladsen

Kvinder	A	B	ialt
med	10	18	28
	y_a	y_b	$y_\cdot$
uden	30	85	115
	$m_a - y_a$	$m_b - y_b$	$m_\cdot - y_\cdot$
ialt	40	103	143
	m_a	m_b	$m_\cdot$

Den statistiske model bliver her tilsvarende

$$(E_k, (P_{(p_{ka},p_{kb})})_{(p_{ka},p_{kb})\in[0,1]^2})$$

hvor udfaldsrummet er givet ved

$$E_k = \{0,1,\cdots,40\} \times \{0,1,\cdots,103\}$$

Fordelingen er dermed også tilsvarende givet ved

$$P(Y_a, Y_b) = \binom{m_a}{y_a}(p_{ka})^{y_a}(1-p_{ka})^{m_a-y_a} \cdot \binom{m_b}{y_b}(p_{kb})^{y_b}(1-p_{kb})^{m_b-y_b}$$

Likelihoodfunktionens største værdi for vores observation bliver så

$$(\hat{p}_{ka}(y_a), \hat{p}_{kb}(y_b)) = \left(\frac{y_a}{m_a}, \frac{y_b}{m_b}\right) = \left(\frac{10}{40}, \frac{18}{103}\right)$$

Indsættes y, k og m

Vi antager nu, at $p_k = p_{ka} = p_{kb}$. Vi opstiller så en nulhypotese

$$H_0 : p_k = p_{ka} = p_{kb} \in [0,1]$$

Under nulhypotesen bliver Likelihoodfunktionen

$$
\begin{aligned}
L(y_a, y_b, p_k) &= \binom{m_a}{y_a}p_k^{y_a}(1-p_k^{m_a-y_a}) \cdot \binom{m_b}{y_b}p_k^{y_b}(1-p_k^{m_b-y_b}) \\
&= \binom{m_a}{y_a}\binom{m_b}{y_b}p_k^{y_\cdot}(1-p_k)^{m_\cdot - y_\cdot}
\end{aligned}
$$

For at finde ML-estimatoren $\hat{P}_{k0}$, skal vi maksimere $L(y_a, y_b, p_k)$

$$\hat{p}_{k0}(y_a, y_b) = \frac{y.}{m.} = \frac{28}{143} \approx 0,1958$$

Nu skal vi se om vi kunne tillade os at antage at $p_{ma} = p_{mb}$. Dette gøres ved at finde kvotienttesteststørrelsen

$$\begin{aligned}
Q(y_a, y_b) &= \frac{L(y_a, y_b, \hat{p}_{k0})}{L(y_a, y_b, \hat{p}_{ka}, \hat{p}_{kb})} \\
&\approx \frac{\left(\frac{10+18}{40+103}\right)^{y.} \left(1 - \frac{10+18}{40+103}\right)^{m.-y.}}{\left(\frac{10}{40}\right)^{y_a} \left(1 - \frac{10}{40}\right)^{m_a - y_a} \left(\frac{18}{85}\right)^{y_b} \left(1 - \frac{18}{85}\right)^{m_b - y_b}} \\
&\approx \frac{(0,1958)^{28}(0,8042)^{115}}{(0,25)^{10}(0,75)^{30}(0,2118)^{18}(0,7882)^{85}} \\
&\approx 0,9988 \\
&\Rightarrow -2 \ln Q \approx 0,0024
\end{aligned}$$

Med en frihedsgrad på 1, giver χ^2-fordelingen os her en værdi mellem $0,1\%$ og $0,01\%$. Derfor må vi her konkludere, at vores hypotese om at det er lige så sandsynligt for en kvinde at få en hudlidelse på de to afdelinger, nok bør genovervejes.

Det næste spørgsmål der ville være relevant er om der er nogen forskel om amn så er en kvinde eller mand med hensyn til at få en hudlidelse. Vi kan nu først opstille tallene i et skema

Hudlidelser	Mænd	Kvinder	ialt
med	11	28	39
	$z_m = (x.)$	$z_k = (y.)$	$z.$
uden	61	115	176
	$s_m - z_m = (n. - x.)$	$s_k - z_k = (m. - y.)$	$s. - z.$
ialt	72	143	215
	$s_m = (n.)$	$s_k = (m.)$	$s.$

For at undersøge om sandsynligheden for at få en hudlidelse er den samme for mænd og kvinder, opstiller vi igen en nullhypotese, hvor vi vil undersøge om man kan antage at

$$H_0 : p_m = p_k = p \in [0, 1]$$

Under denne nulhypotese, bliver Likelihoodfunktionen

$$L(z_m, z_k, p) = \binom{s_m}{z_m} \binom{s_k}{z_k} p^{z.} (1-p)^{s.-z.}$$

For at finde ML-estimatoren $\hat{P}_0$, skal vi maksimere $L(z_m, z_k, p)$

$$\hat{p}_0(z_m, z_k) = \frac{z.}{s.} = \frac{39}{215} \approx 0,1814$$

Nu skal vi se om vi kunne tillade os at antage at $p_k = p_m = p$. Dette gøres ved at finde kvotienttesteststørrelsen

$$Q(z_m, z_k) = \frac{L(z_m, z_k, \hat{p}_0)}{L(z_m, z_k, \hat{p}_k, \hat{p}_m)}$$

$$
\begin{aligned}
&\approx \frac{\left(\frac{11+28}{72+143}\right)^{z_\cdot}\left(1-\frac{11+28}{72+143}\right)^{s_\cdot - z_\cdot}}{\left(\frac{11}{72}\right)^{z_m}\left(1-\frac{11}{72}\right)^{s_m - z_m}\left(\frac{28}{143}\right)^{z_k}\left(1-\frac{28}{143}\right)^{s_k - z_k}} \\
&\approx \frac{(0,1814)^{39}(0,8186)^{176}}{(0,1528)^{11}(0,8472)^{61}(0,1958)^{28}(0,8042)^{115}} \\
&\approx 0,7368 \\
&\Rightarrow -2\ln Q \approx 0,6108
\end{aligned}
$$

Med en frihedsgrad på 1, giver χ^2-fordelingen os her en værdi godt under 5%, hvor vi så kan konkludere, at vi ikke kan afvise, at sandsynlighedenfor at få en hudlidelse er uafhængig af køn.

Opgave 3.16 i Statistik

To steder i samme bæk er opsat en limfælde, og hvert sted har man observeret, hvor mange insekter af en bestemt art, der er blevet fanget i fælden i løbet af en time. Limfælderne er cirkelrunde udspændte stofstykker, der er smurt med lim og opstillet på tværs i bækken. Den ene har en diameter på 5 cm. Den anden har en diametet på 15 cm. I den første fælde er det fanget 12 insekter, i den anden er det fanget 60. Vi skal nu svare på, om det tyder på om insekttætheden er den samme de to steder.

Vi antager at der er tale om en Poissonfordeling. Vi skal da sammenligne to Poissonfordelinger med proportionale parametre. Vi har to observationer, nemlig

$$x_1 = 12, \qquad x_2 = 60$$

hvor arealet på den anden er $3^2 = 9$ gange så stort som det andet. Udfaldsrummet bliver $E = \mathbf{N}_0^2$, og parameterområdet er $\lambda \in [0, \infty[$. Den statistiske model er her

$$(\mathbf{N}_0^2, (P_{\lambda_1,\lambda_2})_{\lambda_1,\lambda_2 \in [0,\infty[^2})$$

hvor fordelingen er givet ved

$$P_{\lambda_1,\lambda_2}(X = x) = \frac{\lambda_1^{x_1}}{x_1!}e^{-\lambda_1} \cdot \frac{\lambda_2^{x_2}}{x_2!}e^{-\lambda_2}$$

Maksimaliseringsestimatoren for en poissonfordeling er givet ved

$$(\hat{\lambda}_1, \hat{\lambda}_2) = (x_1, x_2)$$

Nulhypotese Det vi er interesserede i at undersøge er, om insektintensiteten er den samme på de målte steder. Men da vores målinger ikke er blevet foretaget over det samme areal stykke stof, kan vi ikke umiddelbart sammenligne vores observationer alene. Derfor bliver vores nulhypotese:

$$H_0 : \frac{\lambda_1}{a_1} = \frac{\lambda_2}{a_2} = \alpha \in [0, \infty]$$

hvor a_i angiver arealet af stofstykket. Maksimaliseringsestimatoren under nulhypotesen er givet ved

$$\hat{\alpha} = \frac{x_1 + x_2}{a_1 + a_2} = \frac{12 + 60}{1 + 9} = 7,2$$

(Læg mærke til, at arealet ikke har nogen specifik enhed her, kun at det ene er 9 gange større end det andet. Det skulle ikke betyde noget om vi målte det i cm. mm. km eller vores egen enhed, siden det er proportionaliteten der er i spil) Nu kan vi opstille kvotienttesten for x

$$\begin{aligned} Q(x) &= \left(\frac{(x_1+x_2)a_1}{(a_1+a_2)x_1}\right)^{x_1} \cdot \left(\frac{(x_1+x_2)a_2}{(a_1+a_2)x_2}\right)^{x_2} \\ &= \left(\frac{(12+60)1}{(1+9)12}\right)^{12} \cdot \left(\frac{(12+60)9}{(1+9)60}\right)^{60} \\ &= (0,6)^{12} \cdot (1,08)^{60} \\ &= 0,002177 \cdot 101,257064 \approx 0,2204 \\ &\Rightarrow \quad -2\ln Q \approx 3,0246 \end{aligned}$$

Dimentionen af parameterområdet $\dim(\Theta) = 2$ og $\dim(\Theta_0) = 1$, derfor bliver frihedsgraden 1. Heraf følger af χ^2-fordelingen, at $\epsilon(12,60) \approx 0,005$. På det kan vi konkludere, at siden sandsynligheden for at få en værdi større end den observerede er større end 5%, hvilket betyder, at vi ikke kan afvise at vores hypotese er rigtig.