Antwoorden hertentamen Statistiek
22 augustus 2012

Opgave 1 Zij X_1, X_2, …, X_n onafhankelijke, identiek verdeelde stochasten met cumulatieve dichtheidsfunctie: $F(x|\theta) = \begin{cases} 0 & \text{als } x < 1 \\ 1 - \left(\frac{1}{x}\right)^\theta & \text{als } x \geq 1 \end{cases}$ met $\theta > 2$.

a 2pt) Bepaal de kansdichtheidsfunctie van X_1. ANTWOORD: $f(x|\theta) = \frac{d}{dx}F(x|\theta) = \frac{\theta}{x^{\theta+1}}$ op $(1, \infty)$.

b 6pt) Bepaal de momentschatter van θ. ANTWOORD: $\mathbb{E}(X_1|\theta) = \int_1^\infty x f(x|\theta)dx = \int_1^\infty \frac{\theta}{x^\theta}dx = \frac{\theta}{\theta-1}$. De momentschatter voldoet aan: $\bar{X} = \frac{\hat{\theta}}{\hat{\theta}-1} \Rightarrow \hat{\theta} = \frac{\bar{X}}{\bar{X}-1}$.

c 6pt) Stel dat $n = 50$, dat $\bar{X} = 1.28$ en er 19 punten in het interval (1-1,2) liggen, 17 punten in het interval (1,2-1,4), 8 punten in het interval (1,4-1,6) en er 6 punten in het interval (1,6-∞) liggen. Bepaal de goodness of fit op basis van de momentschatter met een significantieniveau van $\alpha = 0,1$. (Gebruik $\hat{\theta} = 4,5$ als u onderdeel b) niet hebt gedaan.)
ANTWOORD: Gebruik de Pearson χ^2-test. $\hat{\theta} = \frac{\bar{X}}{\bar{X}-1} = \frac{1.28}{1.28-1} \approx 4,57$
$P(X_i \in (1,0-1,2)) = F(1,2) - F(1) = F(1,2) = 1 - \left(\frac{1}{1,2}\right)^{\hat{\theta}} \approx 0.565$
$P(X_i \in (1,2-1,4)) = F(1,4) - F(1,2) = \left(\frac{1}{1,2}\right)^{\hat{\theta}} - \left(\frac{1}{1,4}\right)^{\hat{\theta}} \approx 0,220$
$P(X_i \in (1,4-1,6)) = F(1,6) - F(1,4) = \left(\frac{1}{1,4}\right)^{\hat{\theta}} - \left(\frac{1}{1,6}\right)^{\hat{\theta}} \approx 0,098$
$P(X_i \in (1,6-\infty)) = 1 - F(1,6) = \left(\frac{1}{1,6}\right)^{\hat{\theta}} \approx 0,117$

	1,0-1,2	1,2-1,4	1,4-1,6	1,6-∞
Observed	19	17	8	6
Expected	28,27	10,99	4,91	5,83
Bijdrage aan X^2	3,04	3.29	1,95	0,00

$X^2 = 3,04 + 3,29 + 1,95 + 0,00 = 8,28$. X^2 is bij benadering χ^2 verdeeld met 2 vrijheidsgraden. 8,28 ligt tussen het 97,5% en het 99% percentiel van de χ^2 verdeling met 2 vrijheidsgraden. Bij een significantieniveau van $\alpha = 0,10$ wordt het model dus verworpen.

d 6pt) Bepaal de meest waarschijnlijke schatter (MLE) van θ.
ANTWOORD: $L = \prod_{i=1}^n \frac{\theta}{x_i^{\theta+1}}$. $l(\theta) = \log L = \sum_{i=1}^n \log\theta - (\theta+1)\log x_i = n\log\theta - (\theta+1)\sum_{i=1}^n \log x_i$. Voor de MLE geldt: $0 = \frac{d}{d\theta}l(\theta) = \frac{n}{\theta} - \sum_{i=1}^n \log x_i \Rightarrow \theta_{MLE} = \frac{n}{\sum_{i=1}^n \log x_i}$.

e 6pt) Bepaal een minimaal voldoende statistiek voor θ en bewijs dat deze minimaal is.
ANTWOORD: $f(x_1, x_2, \ldots, x_n|\theta) = \prod_{i=1}^n \frac{\theta}{x_i^{\theta+1}} = \frac{\theta^n}{\left(\prod_{i=1}^n x_i\right)^{\theta+1}} :=$
$g(T(x_1, x_2, \ldots, x_n), \theta)h(x_1, x_2, \ldots, x_n)$ met $T(x_1, x_2, \ldots, x_n) = \prod_{i=1}^n x_i$, $g(t, \theta) = \frac{\theta^n}{t^{\theta+1}}$ en $h(x_1, x_2, \ldots, x_n) = 1$.
Volgens de factorisatiestelling is T een voldoende statistiek.
Gebruik verder de volgende stelling: Zij $T(x_1, x_2, \ldots, x_n) := T(\mathbf{x})$ een functie zodanig dat $\forall \mathbf{x}, \mathbf{y}$ geldt: $\frac{f(\mathbf{x}|\theta)}{f(\mathbf{y}|\theta)}$ onafhankelijk van $\theta \Leftrightarrow T(\mathbf{x}) = T(\mathbf{y})$. Dan is T een minimaal voldoende statistische grootheid voor θ.

Stel $T(\mathbf{x}) = T(\mathbf{y})$. Dan geldt dat $\frac{f(\mathbf{x}|\theta)}{f(\mathbf{y}|\theta)} = \frac{g(T(\mathbf{x}),\theta)}{g(T(\mathbf{y}),\theta)} = 1$, onafhankelijk van θ.

Neem vervolgens aan dat $\frac{f(\mathbf{x}|\theta)}{f(\mathbf{y}|\theta)}$ onafhankelijk is van θ.
$\frac{f(\mathbf{x}|\theta)}{f(\mathbf{y}|\theta)} = \frac{g(T(\mathbf{x}),\theta)h(x_1,x_2,\ldots,x_n)}{g(T(\mathbf{y}),\theta)h(y_1,y_2,\ldots,y_n)} = \frac{g(T(\mathbf{x}),\theta)}{g(T(\mathbf{y}),\theta)} = \left(\frac{T(\mathbf{y})}{T(\mathbf{x})}\right)^{\theta+1}$. Deze uitdrukking is onafhankelijk van θ als $T(\mathbf{x}) = T(\mathbf{y})$ of als $T(\mathbf{y}) = 0$. Dit laatste kan niet, dus $T(\mathbf{x}) = T(\mathbf{y})$. $\square$

f 6pt) Bepaal de Fisher informatie in 1 waarneming. ANTWOORD: $I(\theta) = -\mathbb{E}\left(\frac{\partial^2}{\partial\theta^2}\log\left(f(x|\theta)\right)\right) = -\mathbb{E}\left(\frac{\partial^2}{\partial\theta^2}\log\left(\frac{\theta}{x^{\theta+1}}\right)\right) = -\mathbb{E}\left(\frac{\partial^2}{\partial\theta^2}\left(\log(\theta) - (\theta+1)\log(x)\right)\right) = \frac{1}{\theta^2}$.

g 8pt) Bepaal de Cramér-Rao ondergrens. Bewijs of de momentschatter asymptotisch efficiënt is.
ANTWOORD: Zij T een zuivere schatter van θ. Dan geldt dat $Var(T) \geq \frac{1}{nI(\theta)} = \frac{\theta^2}{n}$.
De momentschatter is asympotisch zuiver. We zijn geïnteresseerd in $Var(\frac{\bar{X}}{\bar{X}-1})$. We weten, wegens de wet van de grote aantallen, dat $\bar{X} \to \mathbb{E}X$ als $n \to \infty$. Zij $\bar{X} = \mathbb{E}X + \epsilon$ met ϵ klein. Dan geldt $\frac{\bar{X}}{\bar{X}-1} = \frac{\mathbb{E}X+\epsilon}{\mathbb{E}X+\epsilon-1} = \frac{\mathbb{E}X}{\mathbb{E}X-1} - \epsilon\frac{1}{(\mathbb{E}X-1)^2} + O(\epsilon^2)$. Dus $Var(\frac{\bar{X}}{\bar{X}-1}) = Var(\frac{\mathbb{E}X}{\mathbb{E}X-1} - \epsilon\frac{1}{(\mathbb{E}X-1)^2} + O(\epsilon^2)) = Var(\epsilon\frac{1}{(\mathbb{E}X-1)^2} + O(\epsilon^2)) = \frac{1}{(\mathbb{E}X-1)^4}Var(\epsilon + O(\epsilon^2))$
Voor grote n geldt, $Var(\frac{\bar{X}}{\bar{X}-1}) \approx \frac{1}{(\mathbb{E}X-1)^4}Var(\bar{X})$. $Var(\bar{X}) = Var(\frac{1}{n}\sum_{i=1}^n X_i) = \frac{1}{n^2}Var(\sum_{i=1}^n X_i) = \frac{1}{n^2}\sum_{i=1}^n Var(X_i) = \frac{1}{n}Var(X_1) = \frac{1}{n}\mathbb{E}(X - \mathbb{E}X)^2 = \frac{1}{n}\int_1^\infty (x - \frac{\theta}{\theta-1})^2\frac{\theta}{x^{\theta+1}}dx = \frac{1}{n}\int_1^\infty \frac{\theta}{x^{\theta-1}} - \frac{2\theta^2}{(\theta-1)x^\theta} + \frac{\theta^3}{(\theta-1)^2x^{\theta+1}}dx = \frac{1}{n}(\frac{\theta}{\theta-2} - \frac{2\theta^2}{(\theta-1)^2} + \frac{\theta^2}{(\theta-1)^2}) = \frac{\theta}{n(\theta-1)^2(\theta-2)}$.
Hieruit volgt dat $Var(\frac{\bar{X}}{\bar{X}-1}) \approx \frac{1}{(\mathbb{E}X-1)^4}\frac{\theta}{n(\theta-1)^2(\theta-2)} = \frac{\theta(\theta-1)^4}{n(\theta-1)^2(\theta-2)} = \frac{\theta(\theta-1)^2}{n(\theta-2)}$. Omdat $\frac{(\theta-1)^2}{\theta(\theta-2)} > 1$ als $\theta > 2$, is $Var(\frac{\bar{X}}{\bar{X}-1}) > \frac{\theta^2}{n}$ en is de momentschatter dus niet asymptotisch efficiënt.

Opgave 2 Zij X_1, X_2, …, X_n onafhankelijke, identiek verdeelde stochasten met kansdichtheidsfunctie $f(x|\theta) = \begin{cases} 0 & \text{als } x < 0 \\ \theta e^{-\theta x} & \text{als } x \geq 0 \end{cases}$, voor $\theta > 0$, en zij x_1, x_2, …, x_n de waarde die de stochasten hebben aangenomen.
Veronderstel als prior kansdichtheid voor θ dat elke waarde van θ groter dan 0 even waarschijnlijk is.

a 3pt) Waarom is dit een oneigenlijke prior? ANTWOORD: Omdat het geen kansdichtheid is. Er is geen constante c zodanig dat $\int_0^\infty c dx = 1$.

b 3pt) Stel dat de prior gelijk is aan $p(\theta) = \begin{cases} 0 & \text{als } \theta \leq 0 \\ \lambda e^{-\lambda\theta} & \text{als } \theta > 0 \end{cases}$, met $\lambda = 10^{-10}$.
Wat is de interpretatie van de prior kansdichtheid voor θ (en wat is de invloed van het feit dat λ klein is)? ANTWOORD: We benaderen de oneigenlijk prior, door een prior die vrijwel vlak is.

c 9pt) Gebruik de oneigenlijk prior om de posterior kansdichtheid van θ te berekenen. Schrijf deze kansdichtheid als de dichtheid van een bekende kansverdeling.
ANTWOORD: $f_{\Theta|X}(\theta) \sim p(\theta)L(\theta|x_1, x_2, \ldots, x_n) = p(\theta)\prod_{i=1}^n \theta e^{-\theta x_i} = p(\theta)\theta^n e^{-\theta\sum_{i=1}^n x_i}$.
Omdat $p(\theta)$ onafhankelijk is van θ stellen we $p(\theta)$ gelijk aan 1. $f_{\Theta|X}(\theta) \sim \theta^n e^{-\theta\sum_{i=1}^n x_i}$ voor $0 \leq \theta \leq N$. $f_{\Theta|X}(\theta)$ moet tot 1 integreren, want het is een kansdichtheid. We zien dat de posterior verdeling van θ gelijk moet zijn aan een $\Gamma(n+1, \sum_{i=1}^n x_i)$-verdeling.

Opgave 3 Een boswachter wil het aantal dassen in een gebied bepalen. Daartoe heeft hij op een dag vallen gezet en de dassen die in de vallen terecht zijn gekomen zijn geoormerkt. In 10 vallen bleek een das te zitten. Korte tijd later zijn er opnieuw vallen geplaatst (op andere plaatsen in het natuurgebied). Dassen in deze vallen werden gevangen gehouden tot er 22 dassen gevangen waren. 6 van de 22 dassen hadden een eerder uitgedeeld oormerk. We

negeren geboorte en sterfte van dassen gedurende het experiment en we veronderstellen dat geoormerkte dassen en dassen zonder oormerk dezelfde kans hebben om gevangen te worden, i.e., we veronderstellen dat het aantal geoormerkte dassen in de populatie van 22 dassen uit een hypergeometrische verdeling komt. Zij $n = 22$ de steekproefgrootte, zij $D = 10$ het aantal geoormerkte dassen, zij N het aantal dassen in het gebied en $d = 6$ is het aantal geoormerkte dassen in de steekproef.

a 7pt) Bepaal de meest waarschijnlijke schatter van N. ANTWOORD: $L(N|d = 6) = \frac{\binom{D}{d}\binom{N-D}{n-d}}{\binom{N}{n}} = \frac{\binom{10}{6}\binom{N-10}{16}}{\binom{N}{22}}$ als $N \geq 26$ en 0 als $N < 26$. De likelihood is maximaal als $10 : 6$ ongeveer gelijk is aan $N - 10 : 16$, oftewel $N = 220/6 \approx 36.7$. Numeriek testen geeft dat het maximum wordt aangenomen in $N = 36$.

b 8pt) Veronderstel als prior kansdichtheid voor N: $p(N = i) = 1/100$ als $21 < i \leq 121$ en 0 anders. Bereken de volgende ratio's van de posterior kansen: $\frac{P_{posterior}(N=25)}{P_{posterior}(N=40)}$, $\frac{P_{posterior}(N=140)}{P_{posterior}(N=40)}$ en $\frac{P_{posterior}(N=33)}{P_{posterior}(N=40)}$.
ANTWOORD: De posteriori kans dat $N = 25$ is 0, omdat $L(N = 25|d = 6) = 0$.
$P_{posteriori}(N = 140) = 0$ omdat de a priori kans dat $N = 140$ gelijk is aan 0.
De eerste twee ratio's zijn daarom 0.

$$\frac{P_{posterior}(N=33)}{P_{posterior}(N=40)} = \frac{L(N=33)}{L(N=40)} = \frac{\frac{\binom{10}{6}\binom{33-10}{16}}{\binom{33}{22}}}{\frac{\binom{10}{6}\binom{40-10}{16}}{\binom{40}{22}}} = \frac{7733}{7830} \approx 0.99$$

Opgave 4 Een hoogleraar wil onderzoeken of het tentamencijfer afhangt van de leeftijd van de kandidaat. Er zijn 5 kandidaten geweest, $(1 \leq i \leq 5)$. De hoogleraar veronderstelt het volgende model voor het tentamencijfer t_i van kandidaat i: $t_i = \beta_0 + \beta_1 a_i + \epsilon_i$ waarbij ϵ_i i.i.d. normaal verdeelde stochasten zijn met gemiddelde 0 en standaarddeviatie σ en a_i de leeftijd van kandidaat i is. De data staan in de tabel hieronder:

individu i	leeftijd a_i	tentamencijfer t_i
1	17	9
2	21	7
3	22	5
4	23	6
5	25	5

De nulhypothese is dat leeftijd geen invloed heeft op het tentamencijfer. De hoogleraar gebruikt als significantieniveau $\alpha = 0,05$.

a 12pt) Toon aan of de nulhypothese verworpen kan worden.
ANTWOORD: De hoogleraar voert een regressie analyse uit en wil testen of $\beta_!$ significant verschillend is van 0. $\bar{x} = 21,6$, $\bar{y} = 6,4$.
$\widehat{\beta_1} = \frac{\sum_{i=1}^{5}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{5}(x_i-\bar{x})^2} = -\frac{91}{176} \approx -0,52$, $\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x} = \frac{773}{44} \approx 17,57$. De regressielijn is dus: $\hat{y} = \frac{773}{44} - \frac{91}{176}a$.
$\frac{\widehat{\beta_1}-\beta_1}{s_{\widehat{\beta_1}}} \sim t_{n-2}$.
$s^2 = \frac{RSS}{5-2} = \frac{\sum_{i=1}^{5}(y_i-(\widehat{\beta_0}-\widehat{\beta_1}x_i))^2}{3} = \frac{105}{176} \approx 0,60$.

$\widehat{Var(\hat{\beta}_1)} = \frac{5s^2}{5\sum_{i=1}^{5}x_i^2-(\sum_{i=1}^{5}x_i)^2} = \frac{5s^2}{5\cdot 2368-11664} \approx 0,017$.
Hieruit volgt dat het 95% betrouwbaarheidsinterval gegeven worden door:

$\hat{\beta}_1 \pm t_3(0,025)\sqrt{\widehat{Var}(\hat{\beta}_1)}$. $t_3(0,025) = 3,182$. Dus het 95% betrouwbaarheidsinterval voor β_1 is: (-0,93;-0,11). De nulhypothese wordt dus verworpen.

b 3pt) Wat kunt u zeggen over het verwachte tentamencijfer van een kandidaat van 65 jaar?
ANTWOORD: Naïef invullen geeft $\hat{y} = \frac{773}{44} - \frac{91}{176}65 \approx -16$. We kunnen concluderen dat het model niet zinvol is voor leeftijden ver verwijderd van de geobserveerde leeftijden.

Opgave 5 Hoe gevoelig een individu is om een mazelen-infectie te krijgen hangt af van de concentratie afweerstoffen in het bloed. In een studiepopulatie van 12 individuen die ooit de mazelen hebben gehad, wil men onderzoeken of 65-plussers een lagere concentratie afweerstoffen hebben dan individuen die jonger zijn dan 65. Als significantieniveau wordt $\alpha = 0,05$ gebruikt. De concentratie afweerstoffen staat in de volgende tabel:

individu	leeftijd	concentratie afweerstoffen (mIU/ml)
1	33	181
2	82	120
3	71	130
4	23	140
5	67	170
6	17	200
7	69	142
8	66	78
9	3	128
10	87	143
11	71	127
12	12	180

a 4pt) Beschrijf het toetsingsprobleem en kies een geschikte toets om de onderzoeksvraag te beantwoorden.
ANTWOORD: H_0: 65-plussers hebben dezelfde distributie van de concentratie afweerstoffen in het bloed als individuen die jonger zijn dan 65. H_1: 65-plussers hebben een andere distributie met een lager gemiddelde. Dit kan getoetst worden met de Wilcoxon rank-sum test of de Mann-Whitney test.

b 11pt) Gebruik deze toets om de onderzoeksvraag te beantwoorden. ANTWOORD

rank	65 plus	concentratie afweerstoffen (mIU/ml)
1	ja	78
2	ja	120
3	ja	127
4	nee	128
5	ja	130
6	nee	140
7	ja	142
8	ja	143
9	ja	170
10	nee	180
11	nee	181
12	nee	200

De som van de rangordes van de 65 plussers is gelijk aan $1+2+3+5+7+8+9 = 35$. De som van de rangordes van de individuen jonger dan 65 jaar is gelijk aan: $4+6+10+11+12 = 43$. Dit is een afwijking van $43 - 5\frac{13}{2} = 43 - 32.5 = 10,5$ van de verwachte waarde onder H_0. De kans, onder de nulhypothese, op een waarde groter of gelijk aan 43 voor de rank-sum van de kleinste groep is gelijk aan de kans, onder de nulhypothese, op een waarde kleiner of gelijk

aan $32.5 - 10.5 = 22$. Dit is groter dan de kritische waarde van 21, zoals vermeld staat in tabel 8 van Rice. (Exact uitrekenen geeft $P(T \geq 43) = 0,053 > \alpha$ onder de nulhypothese.) Er is dus geen reden om de nulhypothese te verwerpen. De conclusie is dat er op basis van deze data geen reden is om aan te nemen dat 65-plussers een lagere concentratie afweerstoffen hebben dan individuen die jonger zijn dan 65.