Antwoorden Tentamen Inleiding Kansrekening en Statistiek (WISB161)
2017-2018
1 februari 2018

Opgave 1 Stel dat twee stochastische variabelen X en Y een gezamenlijke kansfunctie hebben gegeven door Tabel 1.

Tabel 1: Gezamenlijke kansen $P(X = x, Y = y)$

		x		
		0	1	2
y	1	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{16}$
	2	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$
	3	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{16}$

a 4pt) Wat is de gezamenlijke uitkomstenruimte Ω van (X, Y)?
ANTWOORD: $\Omega = \{0, 1, 2\} \times \{1, 2, 3\} =$
$\{(0, 1), (0, 2), (0, 3), (1, 1), (1, 2), (1, 3), (2, 1), (2, 2), (2, 3)\}$.

b 6pt) Zijn X en Y onafhankelijk? Licht je antwoord toe.
ANTWOORD $P(X = 0) = \sum_{i=1}^{3} P(X = 0, Y = i) = \frac{1}{16} + \frac{1}{8} + \frac{1}{16} = \frac{1}{4}$
$P(X = 1) = \sum_{i=1}^{3} P(X = 1, Y = i) = \frac{1}{8} + \frac{1}{4} + \frac{1}{8} = \frac{1}{2}$
$P(X = 2) = \sum_{i=1}^{3} P(X = 2, Y = i) = \frac{1}{16} + \frac{1}{8} + \frac{1}{16} = \frac{1}{4}$
Vanwege symmetrie geldt: $P(Y = i) = P(X = i - 1)$. We zien dat $P(X = i, Y = j) = P(X = i)P(Y = j)$ voor alle $0 \leq i \leq 2$, $1 \leq j \leq 3$, dus X en Y zijn onafhankelijk.

Opgave 2 Stel dat het IQ van studenten die het vak WISB161 volgen normaal verdeeld is met gemiddelde μ en variantie σ^2. Stel dat je wilt testen, met een significantieniveau gelijk aan $\alpha = 0,05$, of het IQ van deze populatie hoger is dan het landelijke gemiddelde van 100. Daartoe meet je het IQ van 5 willekeurig gekozen studenten, en de gevonden IQ-waarden zijn: $\{121, 97, 122, 119, 101\}$.

a 10pt) Voer een geschikte test uit en om te te concluderen of het gemiddelde IQ van deze populatie hoger is dan 100.
Voer een t-test uit. Het steekproefgemiddelde is $\frac{1}{5}(121+97+122+119+101) = \frac{560}{5} = 112$. Een zuivere schatter van de variantie is de steekproefvariantie $s_5^2 = \frac{1}{4}\sum_{i=1}^{5}(x_i - 112)^2 = \frac{1}{4}(9^2+15^2+10^2+7^2+11^2) = \frac{1}{4}(81+225+100+49+121) = \frac{1}{4}(576) = 144$. $\sqrt{5}\frac{\bar{X}_5-\mu}{S_5}$ heeft een student-t-verdeling met 4 vrijheidsgraden. Het 95% -percentiel van deze verdeling is 2.132. Er geldt dus: $P(\sqrt{5}\frac{\bar{X}_5-\mu}{S_5} > 2.132) = 0.05$ en dus $P(\mu < \bar{X}_5 - 2.132\frac{S_5}{\sqrt{5}}) = 0.05$.

Een eenzijdig 95% betrouwbaarheidsinterval voor μ is dus

$(\bar{x}_5 - 2.132\frac{s_5}{\sqrt{5}}; \infty) = (112 - 2.132\frac{12}{\sqrt{5}}; -\infty,)$. Omdat $\sqrt{5} > 2.132$ is de bovengrens van dit interval groter dan 100, oftewel $(100, 558; \infty)$. $\mu = 100$ ligt niet in het 95% betrouwbaarheidsinterval, dus we kunnen op basis van deze steekproef concluderen dat het IQ van studenten die het vak WISB161 volgen hoger is dan het landelijk gemiddelde.

Opgave 3 Zij $-2 \leq \theta \leq 2$ een parameter en zij X_θ een continue stochastische variabele op de reële getallen met kansdichtheidsfunctie

$$f_\theta(x) = \begin{cases} \alpha(\theta) + \theta x & \text{voor } x \in (0,1) \\ 0 & \text{als } x \notin (0,1). \end{cases}$$

Stel dat we een willekeurig steekproef $X_1, X_2, \ldots, X_n$ hebben van grootte n hebben waarbij alle X_i onafhankelijk en identiek verdeeld zijn als X_θ.

a 4pt) Gebruik dat f_θ een kansdichtheid is om te bepalen hoe $\alpha(\theta)$ afhangt van θ.
ANTWOORD: Een kansdichtheidsfunctie moet integreren tot 1, dus we weten: $\int_0^1 f_\theta(x)dx = 1$. Hieruit volgt: $\int_0^1 \alpha(\theta) + \theta x = \alpha(\theta)x + \frac{\theta}{2}x^2|_0^1 = \alpha(\theta) + \frac{\theta}{2} = 1$ en dus $\alpha(\theta) = 1 - \frac{\theta}{2}$

b 2pt) Bepaal $P(X_\theta = \frac{1}{2})$.
ANTWOORD: $P(X_\theta = a) = 0$ voor alle a omdat X_θ een continue stochast is.

c 3pt) Bepaal $E(X_\theta)$.
ANTWOORD: $E(X_\theta) = \int_0^1 x f_\theta(x)dx = \int_0^1 x(1 - \frac{\theta}{2} + \theta x)dx = \frac{1}{2} + \frac{1}{12}\theta$

d 5pt) Bepaal een zuivere schatter van θ op basis van het steekproefgemiddelde.
ANTWOORD Vanwege de i.i.d. geldt: $E(\bar{X}_n) = \frac{1}{2} + \frac{1}{12}\theta$. De methode van momenten geeft als schatter van θ: $\bar{X}_n = \frac{1}{2} + \frac{1}{12}\hat{\theta}$ en dus $\hat{\theta} = 6(2\bar{X}_n - 1)$. Omdat dit een lineaire transformatie van $\bar{X}_n$ is geldt $E(\hat{\theta}) = 6(2E(\bar{X}_n) - 1) = \theta$, dus $\hat{\theta}$ is een zuivere schatter van θ.

e 5pt) Bepaal een formule waaraan de meest waarschijnlijke schatter (maximum likelihood schatter) van θ moet voldoen als we aan mogen nemen dat de absolute waarde van de meest waarschijnlijke schatter kleiner is dan 2.
ANTWOORD: $l(\theta) := \log L(\theta) = \log\left(\prod_{i=1}^n f_\theta(x_i)\right) = \sum_{i=1}^n \log f_\theta(x_i) = \sum_{i=1}^n \log\left(1 - \frac{1}{2}\theta + \theta x_1\right)$. Om het maximum te bepalen differentiëren we naar θ en stellen we de afgeleide gelijk aan nul. $\frac{d}{d\theta}l(\theta)|^{\hat{\theta}} = \sum_{i=1}^n \frac{-\frac{1}{2}+x_i}{1-\frac{1}{2}\hat{\theta}+\hat{\theta}x_i} = 0$.

f 2pt) Stel $n = 2$, $x_1 = \frac{1}{3}$ en $x_2 = \frac{3}{4}$. Bepaal de meest waarschijnlijke schatting van θ.
ANTWOORD $\frac{-\frac{1}{2}+\frac{1}{3}}{1-\frac{1}{2}\hat{\theta}+\hat{\theta}\frac{1}{3}} + \frac{-\frac{1}{2}+\frac{3}{4}}{1-\frac{1}{2}\hat{\theta}+\hat{\theta}\frac{3}{4}} = 0 \Rightarrow -\frac{\frac{1}{6}}{1-\frac{1}{6}\hat{\theta}} + \frac{\frac{1}{4}}{1+\frac{1}{4}\hat{\theta}} = 0 \Rightarrow \hat{\theta} = 1.$

Opgave 4 Stel dat 100 onderzoeksteams elk een data set hebben verzameld en daarmee elk een 90%-betrouwbaarheidsinterval voor een parameter θ geconstrueerd hebben. Veronderstel dat de data sets onafhankelijk zijn.

a 5pt) Bepaal de kansfunctie van het aantal keren dat de daadwerkelijke parameterwaarde θ_0 niet in het 90% betrouwbaarheidsinterval valt.
ANTWOORD: Een 90%-betrouwbaarheidsinterval zal de daadwerkelijke parameter in 90% van de gevallen bevatten en in 10% van de gevallen niet. De gevraagde kansfunctie is dus de kansfunctie een binomiale verdeling. $P(X = k) = \binom{100}{k}(0.1)^k(0.9)^{100-k}$.

b 10pt) Stel dat je twijfelt of elke onderzoeksgroep in staat is om een correct 90%-betrouwbaarheidsinterval te construeren. Je kiest als nulhypothese dat elke onderzoeksgroep correct een 90%-betrouwbaarheidsinterval kan construeren en de alternatieve hypothese is dat niet alle groepen dit kunnen. Om deze test uit te voeren heb je zelf θ_0 met grote nauwkeurigheid bepaalt en op basis hiervan geconcludeerd dat θ_0 bevat was in 80 van de 100 90% betrouwbaarheidsintervallen.

Voer een geschikte test uit met een significantieniveau van $\alpha = 0.0456$. Je mag de normale benadering gebruiken.

ANTWOORD: Onder de nulhypothese is het aantal keren dat θ_0 niet bevat is in een betrouwbaarheidsinterval binomiaal verdeeld met parameters $n = 100$ en $p = 0.1$. Vanwege de dualiteit van hypothesetesten en betrouwbaarheidsintervallen, kunnen we ook een 95,44% betrouwbaarheidsinterval voor p bepalen. De MLE-schatting en de momentschatter van de parameter p is de fractie van de keren dat θ_0 niet in het betrouwbaarheidsinterval viel, i.e., $\hat{p} = 20/100 = 0.2$. Een schatter van de variantie is $\hat{p}(1-\hat{p})$, en die is gelijk aan 0.16. Omdat we een twee-zijdige test hadden, kiezen we nu een symmetrisch betrouwbaarheidsinterval, dus de waarschijnlijkheid dat de daadwerkelijke waarde van p groter dan de bovengrens is is gelijk aan de waarschijnlijkheid dat de daadwerkelijke waarde van p kleiner dan de ondergrens is, en dat is gelijk aan 0.0456/2=0.0228. De kans dat een standaardnormale verdeling groter dan 2 is is precies 0.0228 (volgens de tabel van de standaardnormale verdeling). Een symmetrisch 95,44%-betrouwbaarheidsinterval voor p is dus: $(\hat{p} - 2\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}, \hat{p} + 2\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}) = (0.2 - 2\frac{0.4}{10}, 0.2 - 2\frac{0.4}{10}) = (0.2 - 0.08, 2 + 0.08) = (0.12, 0.28)$. $p = 0.1$ is niet bevat in dit 95,44%-betrouwbaarheidsinterval, dus we verwerpen de nulhypothese.

Opgave 5 Zij $X_1, X_2, \dots X_n$ onafhankelijke, identiek verdeelde Bernoulli stochasten met parameter p. Definieer een run als een maximale hoeveelheid opeenvolgende enen. In het onderstaande realisatie met $n = 12$ zijn er 3 runs.

$$0\quad 0\quad 0\quad \underbrace{1\quad 1}_{\text{run 1}}\quad 0\quad \underbrace{1}_{\text{run 2}}\quad 0\quad 0\quad \underbrace{1\quad 1\quad 1}_{run3}$$

Zij $I_i = \begin{cases} 1 & \text{als een run begint op positie } i \\ 0 & \text{anders} \end{cases}$ voor $1 \leq i \leq n$ en zij $R = \sum_{i=1}^{n} I_i$ het aantal runs.

a 5pt) Bepaal de marginale kansverdeling van $I_1, I_2, \dots, I_n$.
ANTWOORD: $P(I_1 = 1) = p$ en $P(I_1 = 0) = 1 - p$.
Voor $1 < i \leq n$ geldt: $P(I_i = 1) = (1 - p)p$, omdat het $(i - 1)$de element nul moet zijn en het $i$de element één en dus $P(I_i = 0) = 1 - p(1 - p)$.

b 5pt) Bepaal de verwachtingswaarde van R.
ANTWOORD: $E(I_1) = p$. Voor $2 \leq i \leq n$ geldt: $E(I_i) = (1 - p)p$. Daarom geldt: $E(R) = E(\sum_{i=1}^{n} I_i) = \sum_{i=1}^{n} E(I_1) = E(I_1)+(n-1)(1-p)p = p+(n-1)(1-p)p$.

c 5pt) **Bonusopgave** Bepaal de Variantie van R.
ANTWOORD: $Var(R) = Var(\sum_{i=1}^{n} I_i) = \sum_{i=1}^{n} Var(I_i) + 2\sum_{i=2}^{n}\sum_{j=1}^{i-1} Cov(I_i, I_j)$. I_i is een Bernoulli-stochast, dus $Var(I_i) = E(I_i)(1 - E(I_i))$. Dan rest het bepalen van $Cov(I_i, I_j)$. Als $j < i - 1$, dan zijn I_i en I_j onafhankelijk, en dus is $Cov(I_i, I_j) = 0$ in dat geval. We weten dat $I_{i-1}I_i = 0$, want als een run begint op positie $i - 1$, kan er niet ook een run beginnen op positie i en vice versa. $Cov(I_i, I_{i-1}) = E(I_iI_{i-1}) - E(I_i)E(I_{i-1}) = -E(I_i)E(I_{i-1})$.

Dit bekent dat $Var(R) = \sum_{i=1}^{n} Var(I_i) + 2\sum_{i=2}^{n}\sum_{j=1}^{i-1} Cov(I_i, I_j) =$

$Var(I_1) + \sum_{i=2}^{n} Var(I_i) + 2Cov(I_1, I_2) + 2\sum_{i=2}^{n-1} Cov(I_i, I_{i+1}) =$
$p(1 - p) + (n - 1)(1 - p)p(1 - (1 - p)p) - 2p^2(1 - p) + 2(n - 2)(p(1 - p))^2$.

Opgave 6 Zij $X_1,\ X_2,\ \ldots,\ X_n$ onafhankelijke identiek verdeelde stochasten met kansdichtheidsfunctie

$$f_\theta(x) = \begin{cases} \frac{1}{1+\theta}e^{-\frac{x}{1+\theta}} & \text{als } x > 0 \\ 0 & \text{als } x \leq 0 \end{cases}$$

met $1 \leq \theta \leq 2$. Zij T_1 een schatter van θ gedefinieerd door door $T_1(X_1, X_2, \ldots, X_n) = \bar{X}_n - 1$ met $\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$.

a 6pt) Bereken de gemiddelde kwadratische fout van T_1.

ANTWOORD: $E(X_1) = \int_0^\infty dx x f_\theta(x) = \int_0^\infty dx x \frac{1}{1+\theta}e^{-\frac{x}{1+\theta}}$. Substitueer $\lambda = \frac{1}{1+\theta}$. Dan krijgen we: $E(X_1) = \int_0^\infty dx x \lambda e^{-\lambda x} = \frac{1}{\lambda}$ want we hebben nu precies de dichthed van een exponentiële verdeling met parameter λ. Dus geldt vanwege lineariteit van de verwachting dat $E(\bar{X}_n) = \frac{1}{\lambda} = 1 + \theta$. T_1 is dus een zuivere schatter en de gemiddelde kwadratische fout is gelijk aan $Var(T_1)$.

$Var(T_1) = Var(\bar{X}_n) = \frac{1}{n^2}Var(\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}nVar(X_1) = \frac{1}{n}(\theta + 1)^2$ waarbij de laatste stap gebruik maakt dat de variantie van een $Exp(\lambda)$-stochast gelijk is aan $\frac{1}{\lambda^2}$. Dus geldt $MSE(T_1) = \frac{(\theta+1)^2}{n}$.

b 6pt) De schatter T_1 kan schattingen produceren die buiten het interval $[1, 2]$ liggen. We definiëren daarom een tweede schatter T_2 door

$$T_2(X_1, X_2, \ldots, X_n) = \min\{2, \max\{1, T_1(X_1, X_2, \ldots, X_n)\}\}$$

met andere woorden, als de schatting $t_1(x_1, x_2, \ldots, x_n)$ groter is dan 2, dan nemen we 2 als schatting, als de schatting kleiner is dan 1, dan nemen we 1 als schatting. Beargumenteer of de gemiddelde kwadratische fout van T_2 groter, kleiner, of gelijk is aan de gemiddelde kwadratische fout van T_1. Merk op, ook als je vraag a) niet hebt kunnen beantwoorden kun je vraag b beantwoorden

ANTWOORD: $E(T_1 - \theta)^2 = \int_{-\infty}^{\infty} dx (T_1(x) - \theta)^2 f_{T_1}(x) =$
$\int_{-\infty}^{1} dx (T_1(x) - \theta)^2 f_{T_1}(x) + \int_1^2 dx (T_1(x) - \theta)^2 f_{T_1}(x) + \int_2^\infty dx (T_1(x) - \theta)^2 f_{T_1}(x) \geq$
$\int_{-\infty}^{1} dx (1 - \theta)^2 f_{T_1}(x) + \int_1^2 dx (T_1(x) - \theta)^2 f_{T_1}(x) + \int_2^\infty dx (2 - \theta)^2 f_{T_1}(x) =$
$(1 - \theta)^2 \int_{-\infty}^{1} dx f_{T_1}(x) + \int_1^2 dx (T_1(x) - \theta)^2 f_{T_1}(x) + (2 - \theta)^2 \int_2^\infty dx f_{T_1}(x) =$
$(1 - \theta)^2 P(T_1 < 1) + \int_1^2 dx (T_1(x) - \theta)^2 f_{T_1}(x) + (2 - \theta)^2 P(T_1 > 2) = E(T_2 - \theta)^2$

c 2pt) Welke schatter heeft de voorkeur?

ANTWOORD: T_2 is weliswaar geen zuivere schatter, maar heeft wel een kleinere MSE, dus T_2 heeft de voorkeur.

Opgave 7 Zij X uniform verdeeld op $(0,1)$ en Y een Pareto verdeling met parameter $\alpha = 1$, i.e., de kansdichtheidsfunctie van Y wordt gegeven door:

$$f_Y(y) = \begin{cases} \frac{1}{y^2} & \text{als } y \geq 1 \\ 0 & \text{anders} \end{cases}$$

X en Y zijn onafhankelijke stochasten en zij $Z = X^Y$.

a 10pt) Bepaal de cumulatieve verdelingsfunctie $F_Z(z)$.
ANTWOORD: Merk op $F_Z(z) = 0$ als $z \leq 0$ en $F_Z(z) = 1$ als $z > 1$. Kies $0 < z < 1$.

$$F_Z(z) = P(Z \leq z) = P(X^Y \leq z) = P(X \leq Z^{1/Y}) = \int_1^\infty dy f_y(y) \int_0^{z^{1/y}} dx f_x(x) = \int_1^\infty dy f_y(y) \int_0^{z^{1/y}} dx = \int_1^\infty dy f_y(y) z^{1/y}$$

Voer de substitutie $1/y = w$ uit.

$$F_Z(z) = \int_1^0 -dw z^w = \int_0^1 dw z^w = \int_0^1 dw e^{\log(z) w} = \frac{1}{\log z} e^{\log(z) w}\Big|_0^1 = \frac{1}{\log z} z^w\Big|_0^1 = \frac{z-1}{\log z}.$$

Einde.