Inleiding Kansrekening en Statistiek (WISB 161)
Tentamen

Sjoerd Dirksen

28 januari 2021, 13:30-16:30

Dit tentamen bestaat uit 4 vragen, die in totaal 45 punten waard zijn. Als het je niet lukt om een onderdeel van een vraag te beantwoorden, dan is het nog steeds mogelijk om de daarop volgende onderdelen te beantwoorden. Schrijf op elk ingeleverd blad jouw naam en studentnummer en nummer de pagina's. Voorzie elk antwoord van een zorgvuldige motivatie. Vereenvoudig jouw antwoorden voor zover mogelijk. Je mag gebruik maken van de tabel aan het einde van het tentamen.

Vraag 1 [9 punten]

Zij (X,Y) een discrete kansvector die waarden aanneemt in $\{1,2,3\} \times \{1,2,3\}$. De waarden van de kansfunctie van (X,Y) worden gegeven in de onderstaande tabel:

			X	
		1	2	3
	1	$*$	$*$	$*$
Y	2	$*$	0	$*$
	3	0	$*$	0

In het bijzonder is $\mathbb{P}(X=1, Y=3) = 0$. Stel nu dat

$$\mathbb{P}(Y=1 \mid X=i) = \frac{2}{3} \text{ en } \mathbb{P}(X=i \mid Y=1) = \frac{i}{6} \text{ voor } i = 1,2,3.$$

Bepaal de missende waarden, gemarkeerd met een $*$, in de tabel.

Uitwerking:

Met de wet van de totale kans vinden we

$$\begin{aligned}
&\mathbb{P}(Y=1) \\
&\quad = \mathbb{P}(Y=1|X=1)\mathbb{P}(X=1) + \mathbb{P}(Y=1|X=2)\mathbb{P}(X=2) + \mathbb{P}(Y=1|X=3)\mathbb{P}(X=3) \\
&\quad = \frac{2}{3}(\mathbb{P}(X=1) + \mathbb{P}(X=2) + \mathbb{P}(X=3)) = \frac{2}{3}.
\end{aligned}$$

Vervolgens vinden we

$$\mathbb{P}(X=i, Y=1) = \mathbb{P}(X=i|Y=1)\mathbb{P}(Y=1) = \frac{i}{9}$$

en

$$\frac{2}{3}\mathbb{P}(X=i) = \mathbb{P}(Y=1|X=i)\mathbb{P}(X=i) = \mathbb{P}(X=i, Y=1) = \frac{i}{9}$$

ofwel

$$\mathbb{P}(X = i) = \frac{i}{6}.$$

Vervolgens vinden we

$$\mathbb{P}(X = 1, Y = 2) = \mathbb{P}(X = 1) - \mathbb{P}(X = 1, Y = 1) = \frac{1}{6} - \frac{1}{9} = \frac{1}{18}$$
$$\mathbb{P}(X = 2, Y = 3) = \mathbb{P}(X = 2) - \mathbb{P}(X = 2, Y = 1) = \frac{1}{3} - \frac{2}{9} = \frac{1}{9}$$
$$\mathbb{P}(X = 3, Y = 2) = \mathbb{P}(X = 3) - \mathbb{P}(X = 3, Y = 1) = \frac{1}{2} - \frac{1}{3} = \frac{1}{6}.$$

De ingevulde tabel is

			X	
		1	2	3
	1	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{1}{3}$
Y	2	$\frac{1}{18}$	0	$\frac{1}{6}$
	3	0	$\frac{1}{9}$	0

Vraag 2 [8 punten]

Zij Y_n binomiaal verdeeld met parameters n en $\frac{3}{4}$. Laat met behulp van de centrale limietstelling zien dat

$$\lim_{n\to\infty} \mathbb{P}(4Y_n \leq 3n) = \frac{1}{2}.$$

Uitwerking:

Als $X_1, \ldots, X_n$ onafhankelijk en Bernoulli verdeeld zijn met parameter $\frac{3}{4}$, dan heeft $\sum_{i=1}^n X_i$ dezelfde verdeling als Y_n. Merk op dat $\mu := \mathbb{E}(X_1) = \frac{3}{4}$ en $\sigma^2 := \mathrm{Var}(X_1) = \frac{3}{4}(1 - \frac{3}{4}) = \frac{3}{16}$. Met deze notatie kunnen we

$$\mathbb{P}(4Y_n \leq 3n) = \mathbb{P}\Big(\sum_{i=1}^n X_i \leq \frac{3n}{4}\Big) = \mathbb{P}\Big(\sum_{i=1}^n X_i - n\mu \leq 0\Big) = \mathbb{P}\Big(\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \leq 0\Big),$$

schrijven. Uit de centrale limiet stelling volgt dat

$$\lim_{n\to\infty} \mathbb{P}\Big(\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \leq 0\Big) = \mathbb{P}(Z \leq 0)$$

waarbij $Z \sim \mathrm{N}(0, 1)$. Aangezien de verdeling van Z continu is en symmetrisch is om 0 geldt

$$1 = \mathbb{P}(Z \geq 0) + \mathbb{P}(Z < 0) = \mathbb{P}(Z \leq 0) + \mathbb{P}(Z < 0) = 2\mathbb{P}(Z \leq 0)$$

ofwel $\mathbb{P}(Z \leq 0) = \frac{1}{2}$.

Vraag 3 [16 punten]

Docent Sieuwerd geeft online het vak *Inleiding Kansrekening en Statistiek* voor Bedrijfskunde. Hij is geïnteresseerd in de kans p^* dat een doorsnee student een meerkeuze quizvraag correct beantwoordt. In het bijzonder wil hij toetsen of de studenten iets opgestoken hebben en beter presteren dan door simpelweg te raden. Hij wil daarom tijdens de hoorcolleges via *Mentimeter* n quizzes afnemen met steeds vier mogelijke antwoorden. Hij noteert bij elke quiz hoeveel foute antwoorden gegeven worden voordat voor het eerst het goede antwoord gekozen wordt. Laat X_i het aantal foute antwoorden in quiz i zijn voordat het eerste goede antwoord gegeven wordt. Sieuwerd gaat er van uit dat de antwoorden die de studenten geven onafhankelijk van elkaar zijn en dat de uitkomsten van verschillende quizzes ook onafhankelijk van elkaar zijn. Bovendien is hij een beetje ijdel en neemt daarom aan dat er oneindig veel studenten in zijn virtuele collegezaal zitten.

(a) Laat zien dat, voor een gegeven $1 \leq i \leq n$, $X_i + 1$ geometrisch verdeeld is met parameter p^*.

Opmerking: in de opgaven hebben we gezien dat

$$\frac{1}{\frac{1}{n}\sum_{i=1}^{n}(X_i + 1)}$$

een consistente schatter van p^* is. Deze uitspraak mag in het vervolg zonder bewijs gebruikt worden.

(b) Sieuwerd bekijkt het toetscriterium

$$\text{verwerp } H_0 \text{ als } \sum_{i=1}^{n}(X_i + 1) \leq tn, \tag{1}$$

voor een $t < 4$. Leg uit waarom (1) als toetscriterium gebruikt kan worden voor de hypothesetoets

$$H_0 : p^* = \frac{1}{4} \qquad \text{v.s.} \qquad H_1 : p^* = \frac{3}{4}.$$

(c) Laat zien dat de kans op een type I fout naar 0 convergeert als $n \to \infty$.

(d) Zij $t = 2$ en neem aan dat Sieuwerd 100 quizzes afneemt. Gebruik Chebyshev's ongelijkheid om te laten zien dat de kans op een type II fout niet groter is dan $\frac{1}{100}$.

(e) De onderwijscoördinator verlost Sieuwerd van zijn ijdelheid door hem er op te wijzen dat er maar 200 studenten voor zijn vak aangemeld zijn. Welke kansfunctie zou Sieuwerd voor X_1 kunnen gebruiken? Druk jouw antwoord uit in termen van p^*.

Uitwerking:

(a) Merk op dat $Y_i = X_i + 1$ de eerste keer is dat het goede antwoord in quiz i wordt gekozen. We kunnen het geven van een antwoord op deze quizvraag door een student zien als een succes/mislukking experiment met kans op succes gelijk aan p^*. We nemen aan dat de antwoorden van de studenten onafhankelijk van elkaar zijn en dat een onbeperkt aantal keer een antwoord gegeven kan worden. Hieruit volgt dat Y_i geometrisch verdeeld is met parameter p^*.

(b) Uit de opmerking bij het vorige onderdeel weten we dat $1/\bar{Y}_n$ een consistente schatter van p^* is. Een 'grote' waarde van $1/\bar{Y}_n$ (oftewel een 'kleine' waarde van $\bar{Y}_n$) is daarom een indicatie dat H_1 in plaats van H_0 waar zou kunnen zijn.

(c) De kans op een type I fout is de kans dat H_0 wordt verworpen terwijl H_0 waar is, d.w.z.,

$$\mathbb{P}_{H_0}\Big(\sum_{i=1}^{n}(X_i+1)\le tn\Big)=\mathbb{P}_{H_0}\Big(\frac{1}{\bar{Y}_n}\ge\frac{1}{t}\Big).$$

Onder H_0 geldt $p^*=\frac{1}{4}$ en $1/\bar{Y}_n$ is een consistente schatter van p^*, dus geldt voor elke $\varepsilon>0$

$$\lim_{n\to\infty}\mathbb{P}_{H_0}\Big(\Big|\frac{1}{\bar{Y}_n}-\frac{1}{4}\Big|\ge\varepsilon\Big)=0.$$

In het bijzonder geldt

$$\begin{aligned}\lim_{n\to\infty}\mathbb{P}_{H_0}\Big(\frac{1}{\bar{Y}_n}\ge\frac{1}{t}\Big)&=\lim_{n\to\infty}\mathbb{P}_{H_0}\Big(\frac{1}{\bar{Y}_n}-\frac{1}{4}\ge\frac{1}{t}-\frac{1}{4}\Big)\\&\le\lim_{n\to\infty}\mathbb{P}_{H_0}\Big(\Big|\frac{1}{\bar{Y}_n}-\frac{1}{4}\Big|\ge\frac{1}{t}-\frac{1}{4}\Big)=0\end{aligned}$$

aangezien $\frac{1}{t}-\frac{1}{4}>0$ als $t<4$.

(d) Schrijf $Z=\sum_{i=1}^{n}(X_i+1)=\sum_{i=1}^{n}Y_i$. De kans op een type II fout is de kans dat H_0 niet verworpen wordt terwijl H_1 waar is, d.w.z.,

$$\mathbb{P}_{H_1}\Big(\sum_{i=1}^{n}(X_i+1)>tn\Big)=\mathbb{P}_{H_1}(Z>tn).$$

Onder H_1 is $p^*=\frac{3}{4}$ en dus volgt uit de lineariteit van de verwachtingswaarde

$$\mathbb{E}_{H_1}(Z)=\sum_{i=1}^{n}\mathbb{E}_{H_1}(Y_i)=n\mathbb{E}_{H_1}(Y_1)=\frac{n}{p^*}=\frac{4n}{3}$$

en, aangezien de Y_i onafhankelijk zijn,

$$\mathrm{Var}_{H_1}(Z)=\sum_{i=1}^{n}\mathrm{Var}_{H_1}(Y_i)=n\frac{1-p^*}{(p^*)^2}=\frac{4n}{9}.$$

Met behulp van Chebyshev's ongelijkheid vinden we nu

$$\begin{aligned}\mathbb{P}_{H_1}(Z>tn)&=\mathbb{P}_{H_1}\Big(Z-\mathbb{E}(Z)>tn-\mathbb{E}(Z)\Big)\\&\le\mathbb{P}_{H_1}\Big(|Z-\mathbb{E}(Z)|>tn-\mathbb{E}(Z)\Big)\\&\le\frac{\mathrm{Var}\,Z}{(tn-\mathbb{E}(Z))^2}=\frac{4}{9n(t-\frac{4}{3})^2}.\end{aligned}$$

Als $n=100$ en $t=2$ is de rechterzijde gelijk aan $\frac{1}{n}=\frac{1}{100}$.

(e) We nemen aan dat de studenten braaf zijn en allemaal naar het college komen. Net als in het geval dat er oneindig veel studenten in de zaal zitten geldt

$$\mathbb{P}(X_1=k)=\mathbb{P}(Y_1=k+1)=p^*(1-p^*)^k$$

voor $k=0,\dots,199$. Naast de mogelijkheden dat student $1,\dots,200$ (voor het eerst) het goede antwoord geven, is nu de enige overgebleven mogelijkheid dat geen van de studenten het goede antwoord geeft. De kans op deze gebeurtenis is

$$\mathbb{P}(X_1=200)=1-\sum_{k=0}^{199}\mathbb{P}(X_1=k)=1-\sum_{\ell=1}^{200}\mathbb{P}(Y_1=\ell)=1-\sum_{\ell=1}^{200}p^*(1-p^*)^{\ell-1}.$$

Uit het dictaat weten we dat

$$\sum_{\ell=1}^{\infty} p^*(1-p^*)^{\ell-1} = 1$$

en dus vinden we

$$\begin{aligned}\mathbb{P}(X_1 = 200) &= \sum_{\ell=201}^{\infty} p^*(1-p^*)^{\ell-1} \\ &= p^*(1-p^*)^{200} \sum_{\ell=0}^{\infty}(1-p^*)^{\ell} = \frac{p^*(1-p^*)^{200}}{p^*} = (1-p^*)^{200}.\end{aligned}$$

Vraag 4 [12 punten]

Sifan is een fanatieke hardloopster. Zij gaat elke week precies een uur hardlopen en probeert in dit uur zo ver mogelijk te lopen. Haar vriendin Josefine volgt het vak *Inleiding Kansrekening en Statistiek* en probeert te modelleren hoe vaak Sifan haar persoonlijk record in de toekomst zal kunnen verbeteren. Zij modelleert de afstanden die Sifan in n (toekomstige) weken loopt als een continue kansvector $(W_1, \ldots, W_n)$ en zij neemt aan dat de afstanden $W_1, \ldots, W_n$ onafhankelijk en identiek verdeeld zijn. Josefine negeert de afstanden die Sifan in het verleden heeft gelopen. Merk op dat Sifan een nieuw persoonlijk record vestigt in week k als

$$W_k \geq \max\{W_1, \ldots, W_{k-1}\}.$$

(a) Laat eerst zien dat

$$\mathbb{P}(Y \leq Z) = \int_{-\infty}^{\infty} F_Y(z) f_Z(z) dz$$

als (Y, Z) een continue kansvector is en Y en Z onafhankelijk zijn.

(b) Laat met behulp van deel (a) zien dat de kans dat Sifan in week n een nieuw persoonlijk record vestigt gelijk is aan $\frac{1}{n}$.

(c) Laat zien dat het verwachte aantal nieuwe persoonlijk records dat tot en met week n gevestigd wordt gelijk is aan $\sum_{k=1}^{n} \frac{1}{k}$.

Opmerking: het is bekend dat $\lim_{n\to\infty} \sum_{k=1}^{n} \frac{1}{k} = \infty$.

Uitwerking:

(a) Als (Y, Z) een continue kansvector is en Y en Z zijn onafhankelijk, dan geldt

$$f_{Y,Z}(y, z) = f_Y(y) f_Z(z), \qquad \text{voor alle } y, z \in \mathbb{R}.$$

Met behulp van stellingen uit het dictaat vinden we

$$\begin{aligned}\mathbb{P}(Y \leq Z) &= \mathbb{E}(1_{\{Y \leq Z\}}) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} 1_{\{y \leq z\}}(y, z) f_{Y,Z}(y, z)\ dydz \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} 1_{\{y \leq z\}}(y, z) f_Y(y) f_Z(z)\ dydz\end{aligned}$$

$$
\begin{aligned}
&= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} 1_{\{y \le z\}}(y, z) f_Y(y)\ dy \right] f_Z(z) dz \\
&= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{z} f_Y(y)\ dy \right] f_Z(z) dz \\
&= \int_{-\infty}^{\infty} F_Y(z) f_Z(z) dz.
\end{aligned}
$$

(b) Schrijf $W := W_1$. Aangezien $W_1, \ldots, W_n$ identiek verdeeld zijn geldt

$$
F_{W_i} = F_W \text{ en } f_{W_i} = f_W, \qquad \text{voor alle } i = 1, \ldots, n.
$$

Zij $Y := \max\{W_1, \ldots, W_{n-1}\}$. Aangezien $W_1, \ldots, W_{n-1}$ onafhankelijk zijn, volgt uit een stelling uit het dictaat dat

$$
F_Y(y) = \prod_{i=1}^{n-1} F_{W_i}(y) = (F_W(y))^{n-1}.
$$

Aangezien $W_1, \ldots, W_n$ onafhankelijk zijn, zijn Y en $Z := W_n$ ook onafhankelijk. Met behulp van deel (a) vinden we nu

$$
\begin{aligned}
&\mathbb{P}(\max\{W_1, \ldots, W_{n-1}\} \le W_n) \\
&= \int_{-\infty}^{\infty} \mathbb{P}(\max\{W_1, \ldots, W_{n-1}\} \le w)\ f_{W_n}(w) dw \\
&= \int_{-\infty}^{\infty} (F_W(w))^{n-1}\ f_W(w) dw \\
&= \lim_{a \to -\infty} \lim_{b \to \infty} \int_a^b (F_W(w))^{n-1}\ f_W(w) dw \\
&= \lim_{a \to -\infty} \lim_{b \to \infty} \frac{1}{n} (F_W(w))^n \Big|_{w=a}^{w=b} = \frac{1}{n}.
\end{aligned}
$$

Hier hebben we gebruikt dat F_W een verdelingsfunctie is en dus geldt

$$
\lim_{b \to \infty} F_W(b) = 1, \qquad \lim_{a \to -\infty} F_W(a) = 0.
$$

(c) Zij Y_k de kansvariabele gedefinieerd door

$$
Y_k = \begin{cases} 1 & \text{als } W_k \ge \max\{W_1, \ldots, W_{k-1}\} \\ 0 & \text{anders.} \end{cases}
$$

Dan is het aantal persoonlijk records Z_n dat gevestigd wordt tot en met week n gelijk aan $Z_n = \sum_{k=1}^n Y_k$ en uit deel (a) weten we dat Y_k Bernoulli verdeeld is met succeskans $\frac{1}{k}$. Uit de lineariteit van de verwachtingswaarde volgt nu

$$
\mathbb{E}(Z_n) = \sum_{k=1}^n \mathbb{E}(Y_k) = \sum_{k=1}^n \frac{1}{k}.
$$

Een aantal belangrijke kansverdelingen

Discrete kansverdelingen

Naam	Kansfunctie	Verwachtingswaarde	Variantie
Ber(p)	$\mathbb{P}(X=1)=p=1-\mathbb{P}(X=0)$	p	$p(1-p)$
Binom(n,p)	$\mathbb{P}(X=k)=\binom{n}{k}p^k(1-p)^{n-k},\ \ k=0,\dots,n$	np	$np(1-p)$
Geom(p)	$\mathbb{P}(X=k)=p(1-p)^{k-1},\ \ k\in\mathbb{N}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Pois(λ)	$\mathbb{P}(X=k)=e^{-\lambda}\frac{\lambda^k}{k!},\ \ k\in\mathbb{N}\cup\{0\}$	λ	λ

Continue verdelingen

Naam	Kansdichtheid	Verwachtingswaarde	Variantie
Unif(a,b)	$f(x)=\begin{cases}\frac{1}{b-a} & \text{als } a<x<b\\ 0 & \text{elders}\end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exp(λ)	$f(x)=\begin{cases}\lambda e^{-\lambda x} & \text{als } x>0\\ 0 & \text{elders}\end{cases}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
N(μ,σ^2)	$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-(x-\mu)^2/(2\sigma^2))$	μ	σ^2