

Statistiek voor A.I.

College 4

Donderdag 20 September 2012

2 Deductieve statistiek

Kansrekening

THE SCIENCE NEWS CYCLE

JORGE CHAM © 2009

Start Here



YOUR GRANDMA

...eventually making it to...



4 LOCAL EYEWITNESS NEWS

...and caught on ...

Cable NEWS



We saw it on a Blog!

A causes B all the time
What will this mean for Obama?
BREAKING NEWS BREAKING NEWS BREA

...then noticed by...



THE INTERNETS



...is translated by...

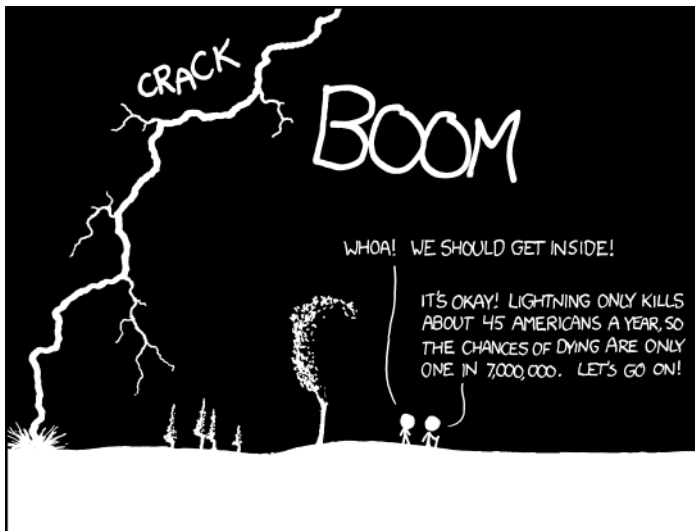


...which is then picked up by...



...who are read by ...

Context



THE ANNUAL DEATH RATE AMONG PEOPLE
WHO KNOW THAT STATISTIC IS ONE IN SIX.

Vandaag:

- Eigenschappen kansen
- Oneindige discrete uitkomstenruimtes
- Continue uitkomstenruimtes
- Continue stochasten

Definitie Gegeven een uitkomstenruimte S is een *stochast* (*stochastische variabele*, *random variable*) een functie die aan elk punt in S een numerieke waarde (een waarde in $\mathbb{R}$) toevoegt. Stochasten worden aangeduid met X, Y, Z . $\text{rng}(X)$ is het beeld van X : $\text{rng}(X) = \{X(s) \mid s \in S\}$.

Als S aftelbaar is dan heet X *discreet* en is de *kansverdeling* (*verdeling*, *distributie*, *distributiefunctie*) van X de functie f (m in boek) gegeven door:

$$f(i) = P(X = i) = P(\{s \in S \mid X(s) = i\}).$$

Er geldt:

- $f : \text{rng}(X) \rightarrow [0, 1]$,
- $\sum_{i \in \text{rng}(X)} f(i) = \sum_{i \in \text{rng}(X)} P(\{s \in S \mid X(s) = i\}) = \sum_{s \in S} P(s) = 1$,
- Voor elke gebeurtenis $A \subseteq \text{rng}(X)$:

$$P(A) = \sum_{i \in A} f(i) = \sum_{\{s \in S \mid X(s) \in A\}} P(s).$$

Antwoord op een vraag: Simpsons paradox

Bij een zekere universiteit solliciteren gedurende een zekere maand 40 mannen en 40 vrouwen.

	aangenomen	afgewezen	percentage
♀	20	20	50
♂	24	16	60

Men spreekt er schande van.

Vervolgens verdeelt men de sollicitaties naar faculteit.

Faculteit Bètawetenschappen:

	aangenomen	afgewezen	percentage
♀	12	18	40
♂	3	7	30

Faculteit Geesteswetenschappen:

	aangenomen	afgewezen	percentage
♀	8	2	80
♂	21	9	70

Men spreekt zijn goedkeuring uit.

Eigenschappen van kansen

Stelling $P(A) = P(A \cap B) + P(A \cap \overline{B})$.

Definitie $B_1, \dots, B_n$ vormen een *partitie* van de uitkomstenruimte als zij onderling disjunct zijn en $S = B_1 \cup B_2 \cup \dots \cup B_n$.

Stelling Als $B_1, \dots, B_n$ een partitie van de uitkomstenruimte is, dan geldt

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + \dots + P(A \cap B_n).$$

Eigenschappen van kansen: partitie

Voorbeeld Het percentage mensen met dyslexie is 20%, met dyscalculi 4% en 0.9% heeft dyslexie én dyscalculi.

Wat is het percentage mensen dat aan precies één van de twee lijdt?

$$0.2 = P(\text{dyslexie}) = P(\text{dyslexie en dyscalculi}) + P(\text{dyslexie en niet dyscalculi})$$

$$0.04 = P(\text{dyscalculi}) = P(\text{dyslexie en dyscalculi}) + P(\text{dyscalculi en niet dyslexie}).$$

Dus

$$P(\text{dyslexie en niet dyscalculi}) = 0.2 - 0.009 = 0.191.$$

$$P(\text{dyscalculi en niet dyslexie}) = 0.04 - 0.009 = 0.031.$$

Dat geeft

$$P(\text{dyslexie en niet dyscalculi}) + P(\text{dyscalculi en niet dyslexie}) = 0.191 + 0.031 = 0.222.$$

Eigenschappen van kansen: samenvatting

- $0 \leq P(A) \leq 1$ en $P(S) = 1$ (axioma's).
- $P(\bar{A}) = 1 - P(A)$.
- Als $A_1, A_2, \dots$ onderling disjunct zijn, dan geldt (axioma)

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = P(A_1) + P(A_2) + \dots$$

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
- $P(A \cup B \cup C) =$
 $P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$.
- $P(A) = P(A \cap B) + P(A \cap \bar{B})$.
- Als $B_1, \dots, B_n$ een partitie van de uitkomstenruimte is, dan geldt

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + \dots + P(A \cap B_n).$$

Oneindige discrete uitkomstenruimtes

Voorbeeld Een dobbelsteen wordt gegooit totdat er de eerste keer een 2 wordt gegooit. De stochast X is het aantal maal dat er gegooit moet worden. X heeft dus oneindig veel mogelijke waarden en de bijbehorende uitkomstenruimte is oneindig: $S = \{1, 2, 3, 4, \dots\} = \mathbb{N}_{>0}$.

$$P(X = 1) = \frac{1}{6} \quad P(X = 2) = \frac{5}{6} \cdot \frac{1}{6} = \frac{5}{36}$$

$$P(X = 3) = \frac{5}{6} \cdot \frac{5}{6} \cdot \frac{1}{6} = \frac{25}{216} \quad \dots$$

$$P(X = n) = \frac{5^{n-1}}{6^n}$$

Er moet $P(S) = 1$ gelden, en inderdaad:

$$\sum_{n=1}^{\infty} P(X = n) = \frac{1}{6} + \frac{5}{36} + \frac{25}{216} + \dots = 1$$

Antwoord op een vraag: natuurlijke getallen

Voorbeeld Wat is de kans dat je uit een zak gevuld met de natuurlijke getallen een 7 trekt?

Intuïtie: voor elk getal is de kans om het te trekken even groot.

Stel dat die kans p is, en laat X de stochast zijn. Dan is voor elk getal n , $P(X = n) = p$.

Omdat de gebeurtenissen disjunct zijn en $P(X \in \mathbb{N}) = 1$ geldt

$$1 = P(X \in \mathbb{N}) = \sum_{n=1}^{\infty} P(X = n) = \sum_{n=1}^{\infty} p = \begin{cases} 0 & \text{als } p = 0 \\ \infty & \text{als } p > 0. \end{cases}$$

Een tegenspraak: $1 = 0$ of $1 = \infty$.

Conclusie: aan dit experiment kan geen kansverdeling toegekend worden die overeenkomt met bovengenoemde intuïtie. Aan de vraag wat $P(X = 7)$ is, kan daarom geen zinvolle betekenis gegeven worden.

Let wel: de eis van uniformiteit van de verdeling is hier essentieel, want de vorige slide toonde een voorbeeld van een (niet-uniforme) kansverdeling op een discrete oneindige uitkomstenruimte.

Continue uitkomstenruimtes

Continue uitkomstenruimtes: kansen toekennen

Intuïtie:



Als X willekeurige waarden aanneemt in $[0, 1]$, dan moet gelden dat

$$P(X \in [0, r]) = P(X \in [1-r, 1]).$$



Als X willekeurige waarden aanneemt in $[r, s]$ en $(d - c) = 2(b - a)$, dan moet gelden dat

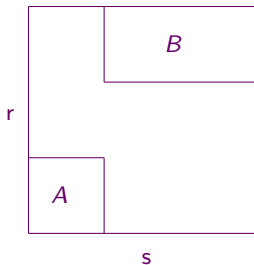
$$P(X \in [c, d]) = 2 \cdot P(X \in [a, b]).$$

Hieraan wordt voldaan als

$$P(X \in [a, b]) = \frac{(b - a)}{(s - r)}.$$

Continue uitkomstenruimtes: kansen toekennen

Intuïtie:



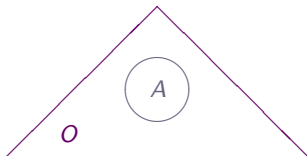
Als X willekeurige waarden aanneemt in de rechthoek en de oppervlakte van B is twee maal de oppervlakte van A , dan moet gelden dat

$$P(X \in B) = 2 \cdot P(X \in A).$$

Hieraan wordt voldaan als

$$P(X \in A) = \frac{\text{oppervlakte } A}{\text{oppervlakte rechthoek}} = \frac{\text{oppervlakte } A}{r \cdot s}.$$

Intuïtie:



Als X willekeurige waarden aanneemt in een oppervlak $O \subseteq \mathbb{R}^2$ en A is een oppervlak in O , dan moet gelden dat

$$P(X \in A) = \frac{\text{oppervlakte } A}{\text{oppervlakte } O}.$$

Continue stochasten

Continue stochasten

Intuïtie:

Discrete stochasten:

$$P(X \in A) = \sum_{i \in A} P(X = i).$$

Continue stochasten:

$$\sum_{i \in A} \sim \int_A.$$

Definitie Een stochast X op uitkomstenruimte S is *continu* als er een *kansdichtheid* $f : \text{rng}(X) \rightarrow \mathbb{R}_{\geq 0}$ is zodat voor elke $A \subseteq \text{rng}(X)$ geldt:

$$P(X \in A) = \int_A f(x) dx \quad \text{en} \quad \int_{\text{rng}(X)} f(x) dx = 1.$$

Vergelijk

Voor een discrete stochast X met kansverdeling f geldt $f : S \rightarrow [0, 1]$ en

$$P(X \in A) = \sum_{i \in A} f(i) \quad \text{en} \quad \sum_{i \in \text{rng}(X)} f(i) = 1.$$

Conventie: Vaak wordt $P(A)$ i.p.v. $P(X \in A)$ geschreven.

Voorbeeld Zij X de continue stochast bij het willekeurig kiezen van een punt in $[0,1]$. Dan heeft X de volgende kansdichtheid:

$$f(x) = \begin{cases} 1 & \text{als } 0 \leq x \leq 1 \\ 0 & \text{anders.} \end{cases}$$

Inderdaad geldt dat

$$\int_{-\infty}^{\infty} f(x) dx = 1.$$

In overeenstemming met de intuïtie geldt voor elk interval $[a, b]$ in $[0,1]$:

$$P(X \in [a, b]) = P(a \leq X \leq b) = \int_a^b f(x) dx = \int_a^b 1 dx = x \Big|_a^b = b - a.$$

Voorbeeld Zij X de continue stochast bij het willekeurig kiezen van een punt in $[r, s]$. Dan heeft X de volgende kansdichtheid:

$$f(x) = \begin{cases} \frac{1}{s-r} & \text{als } r \leq x \leq s \\ 0 & \text{anders.} \end{cases}$$

Inderdaad geldt dat

$$\int_{-\infty}^{\infty} \frac{1}{s-r} dx = 1.$$

In overeenstemming met de intuïtie geldt voor alle $r \leq a \leq b \leq s$ dat

$$P(a \leq X \leq b) = \int_a^b f(x) dx = \int_a^b \frac{1}{s-r} dx = \frac{x}{s-r} \Big|_a^b = \frac{b-a}{s-r}.$$

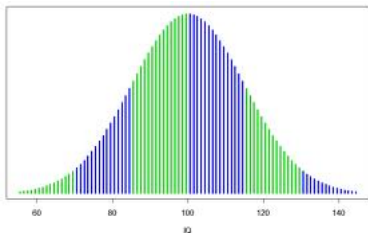
Voorbeeld Zij X de continue stochast bij het willekeurig kiezen van een punt in een rechthoek met lengte r en hoogte s . Dan heeft X de volgende kansdichtheid:

$$f(x, y) = \begin{cases} \frac{1}{r \cdot s} & \text{als } x \text{ in de rechthoek ligt} \\ 0 & \text{anders.} \end{cases}$$

Er geldt dat

$$\begin{aligned} P(X \in [a, b] \times [c, d]) &= \int_c^d \int_a^b f(x, y) dx dy = \int_c^d \int_a^b \frac{1}{r \cdot s} dx dy = \\ &= \int_a^b \frac{(d - c)}{r \cdot s} dx = \frac{(d - c) \cdot x}{r \cdot s} \Big|_a^b = \frac{(b - a)(d - c)}{r \cdot s}. \end{aligned}$$

Voorbeeld De kansdichtheid f van het IQ van mensen:



$$P(85 \leq X \leq 115) = \int_{85}^{115} f(x) = 0.68.$$

$$P(70 \leq X \leq 130) = \int_{70}^{130} f(x) = 0.95.$$

Antwoord op een vraag: telefoontjes

Voorbeeld Zij X de continue stochast waarvan de waarden de tijdsduur (in minuten) tussen twee opeenvolgende telefoontjes in een alarmcentrale zijn. Uit onderzoek blijkt dat de kansdichtheid van X de volgende is:

$$f(x) = \begin{cases} \frac{1}{5}e^{-\frac{x}{5}} & \text{als } x \geq 0 \\ 0 & \text{anders.} \end{cases}$$

Merk op: f neemt willekeurig grote waarden aan.

De kans dat de tijdsduur tussen twee telefoontjes tussen de 6 en 7 minuten ligt:

$$P(6 \leq X \leq 7) = \int_6^7 \frac{1}{5}e^{-\frac{x}{5}} dx = -e^{-\frac{x}{5}} \Big|_6^7 = -e^{-\frac{7}{5}} + e^{-\frac{6}{5}} = 0.25.$$

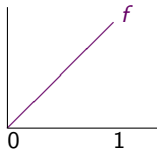
De kans dat de tijdsduur tussen twee telefoontjes groter is dan 5 minuten:

$$P(X > 5) = \int_5^{\infty} \frac{1}{5}e^{-\frac{x}{5}} dx = -e^{-\frac{x}{5}} \Big|_5^{\infty} = e^{-1} = \frac{1}{e} = 0.37.$$

Antwoord op een vraag: magneet

Voorbeeld Een metalen balletje wordt geworpen op een langwerpige magneet van 1 m. waarvan de magnetische kracht gelijkmatig toeneemt naar de rechter kant. X is de continue stochast waarvan de waardes de plaats van het balletje op de magneet zijn. Gegeven is dat de kansdichtheid van X f is, waarbij

$$f(x) = \begin{cases} 2x & \text{als } 0 \leq x \leq 1 \\ 0 & \text{anders.} \end{cases}$$



De kans dat het balletje op de rechter helft van de magneet valt is:

$$P(0.5 \leq X \leq 1) = \int_{0.5}^1 2x dx = x^2 \Big|_{0.5}^1 = 0.75.$$

De kans dat het balletje op de linker helft van de magneet valt is:

$$P(0 \leq X \leq 0.5) = \int_0^{0.5} 2x dx = x^2 \Big|_0^{0.5} = 0.25.$$

Continue stochasten: eigenschappen

De eigenschappen van discrete stochasten die behandeld zijn gelden ook voor continue stochasten:

- $0 \leq P(A) \leq 1$ en $P(S) = 1$.
- $P(\bar{A}) = 1 - P(A)$.
- Als $A_1, A_2, \dots$ onderling disjunct zijn, dan geldt

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = P(A_1) + P(A_2) + \dots$$

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
- $P(A \cup B \cup C) =$
 $P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$.
- $P(A) = P(A \cap B) + P(A \cap \bar{B})$.
- Als $B_1, \dots, B_n$ een partitie van de uitkomstenruimte is, dan geldt

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + \dots + P(A \cap B_n).$$

Continue stochasten: de kans op een getal

In tegenstelling tot discrete stochasten kan bij continue stochasten $f(r)$ niet geïnterpreteerd worden als $P(X = r)$.

Voor continue stochasten is $P(X = r) = 0$ voor elk getal r :

$$\begin{aligned} P(X = r) &= \lim_{\epsilon \rightarrow 0} P(r - \epsilon \leq X \leq r + \epsilon) = \lim_{\epsilon \rightarrow 0} \int_{r-\epsilon}^{r+\epsilon} f(x) dx = \\ &= \lim_{\epsilon \rightarrow 0} F(x) \Big|_{r-\epsilon}^{r+\epsilon} = \lim_{\epsilon \rightarrow 0} (F(r + \epsilon) - F(r - \epsilon)) = 0. \end{aligned}$$

Hierbij is F de primitieve van f (d.w.z. f is de afgeleide van F).

Finis