

Statistiek voor A.I.

College 7

Dinsdag 2 Oktober

2 Deductieve statistiek

Kansrekening

Vraag: test

Een test op *HIV* is 90% betrouwbaar: als een persoon *HIV* heeft is de kans op een positieve uitslag 0.9, en als een persoon geen *HIV* heeft is de kans op een positieve uitslag 0.1. De kans op *HIV* is 0.05.

Een dame ondergaat de test en de uitslag is positief. Wat is de kans dat zij *HIV* heeft?

Vraag: spam

De kans dat een email spam is, is 99%. Een spamfilter is 98% betrouwbaar: 98% van alle spam wordt als spam geclassificeerd en 2% van niet-spam wordt als spam geclassificeerd.

Een email wordt als spam geclassificeerd. Hoeveel is de kans dat het daadwerkelijk spam is toegenomen door deze informatie?

Vandaag :

- Voorwaardelijke kansen
- Stelling van Bayes

Voorwaardelijke kansen

Voorwaardelijke kansen: kinderen

49% van alle geboortes is een meisje. De familie de Bruin heeft vier kinderen. Meneer de Bruin zegt: *wij hebben gelukkig niet allemaal jongens. Wat is de kans dat ik precies drie meisjes heb?*

De volgende dag ga jij bij hen op bezoek, en meneer de Bruin doet de deur open met een dochter aan de hand. Hij zegt: *dit is mijn jongste kind, Marianne. Wat is de kans dat ik precies drie meisjes heb?* Geef jij beide keren hetzelfde antwoord?

X is het aantal meisjes, A is de gebeurtenis dat het jongste kind een meisje is. Gevraagd wordt naar de kansen

$$P(X = 3 | X \geq 1) = \frac{P(X = 3)}{P(X \geq 1)} = \frac{\binom{4}{3}(0.49)^3(0.51)}{1 - (0.51)^4} = 0.26.$$

$$P(X = 3 | A) = \frac{P(X = 3 \cap A)}{P(A)} = \frac{(0.49)\binom{3}{2}(0.49)^2(0.51)}{0.49} = 0.37.$$

Jij geeft niet hetzelfde antwoord.

Voorwaardelijke kansen: eigenschappen

Veel eigenschappen die we eerder zagen gelden ook voor voorwaardelijke kansen.

Stelling Voor alle gebeurtenissen A en B geldt

$$P(A | B) + P(\bar{A} | B) = 1.$$

Stelling Als $A_1, \dots, A_n$ onderling disjunct zijn, dan geldt

$$P\left(\bigcup_{i=1}^n A_i \mid B\right) = \sum_{i=1}^n P(A_i \mid B).$$

Definitie De *prosecutor's fallacy* bestaat uit het hanteren van de kans $P(D | O)$ als de kans op verdachtes onschuld (O) op grond van gevonden bewijsgrond (D), in plaats van de kans $P(O | D)$.

$P(D | O)$ is de kans dat D plaatsvindt gegeven dat de verdachte onschuldig is.

$P(O | D)$ is de kans dat de verdachte onschuldig is gegeven dat D plaatsvindt.

Merk op:

$$P(O | D) = P(D | O) \frac{P(O)}{P(D)}.$$

De prosecutor's fallacy vond geregeld plaats in de rechtspraak.

Voorwaardelijke kansen: prosecutor's fallacy

Voorbeeld Stel dat in een populatie van 10.000 mensen er 2 mensen zijn die aan het DNA type dat bij een misdrijf gevonden is voldoen (inclusief de crimineel).

D is de gebeurtenis dat de een persoon dat DNA type heeft, O de gebeurtenis dat een persoon onschuldig is.

$$P(D | O) = \frac{P(D \cap O)}{P(O)} = \frac{1}{9999}$$

$$P(O | D) = \frac{P(D \cap O)}{P(D)} = \frac{1}{2}.$$

Voorwaardelijke kansen: prosecutor's fallacy

Voorbeeld De zaak *People vs Collins* (1968): een portomonee wordt gestolen en een getuige zegt een blonde vrouw met staart te hebben zien vluchten in een gele auto bestuurd door een zwarte man met baard. Een aantal dagen later wordt er een paar dat aan deze beschrijving voldoet gearresteerd. Een wiskundige berekent dat de kans dat een willekeurig gekozen paar aan deze omschrijving voldoet $8.3 \cdot 10^{-8}$ is. Het opgepakte paar wordt veroordeeld. De Hoge Raad vindt dat er nauwkeuriger gerekend moet worden. Op grond van de getuigenverklaringen bestaat er een paar met de beschreven eigenschappen. Wat is de kans dat er nog zo een paar bestaat?

Stel dat de populatie uit n stellen bestaat en laat p de kans zijn dat een willekeurige paar uit de populatie aan de beschrijving voldoet. A_i : het i^e paar voldoet aan de beschrijving. O : precies 1 paar voldoet aan de beschrijving. A (B): tenminste 1 (2) paren voldoen aan de beschrijving. Merk op: $A = B \cup O$, en O en B zijn disjunct.

$$P(A) = 1 - (1 - p)^n \quad P(O) = np(1 - p)^{n-1}.$$

$$P(B) = P(A) - P(O) = 1 - (1 - p)^n - np(1 - p)^{n-1}.$$

Omdat $B \subseteq A$:

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(B)}{P(A)} = \frac{1 - (1 - p)^n - np(1 - p)^{n-1}}{1 - (1 - p)^n}.$$

De beroving vond plaats in een dicht bevolkt gebied: met $n = 8.000.000$ en $p = 8.3 \cdot 10^{-8}$ geldt $P(B|A) = 0.2966$.

Onafhankelijkheid

Onafhankelijkheid

Definitie Twee gebeurtenissen A en B zijn *onafhankelijk* (*independent*) als

$$P(A|B) = P(A) \text{ en } P(B|A) = P(B).$$

Onafhankelijkheid

Voorbeeld Een dobbelsteen en een munt worden geworpen. De gebeurtenis dat er 6 gegooid wordt is onafhankelijk van de gebeurtenis dat er K gegooid wordt:

$k1$	$k2$	$k3$	$m1$	$m2$	$m3$
$k4$	$k5$	$k6$	$m4$	$m5$	$m6$

$$P(X = 6 | Y = K) = \frac{P(X = 6 \cap Y = K)}{P(Y = K)} = \frac{\frac{1}{12}}{\frac{6}{12}} = \frac{1}{6} = P(X = 6).$$

$$P(Y = K | X = 6) = \frac{P(X = 6 \cap Y = K)}{P(X = 6)} = \frac{\frac{1}{12}}{\frac{2}{12}} = \frac{1}{2} = P(Y = K).$$

Stelling A en B zijn onafhankelijk dan en slechts dan als geldt dat

$$P(A \cap B) = P(A)P(B).$$

Bewijs Per definitie geldt

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Daaruit volgt dat

$$P(A \cap B) = P(A|B)P(B).$$

Als A en B onafhankelijk zijn geldt $P(A|B) = P(A)$. Dit geeft

$$P(A \cap B) = P(A)P(B).$$

Omgekeerd, als $P(A \cap B) = P(A)P(B)$, dan volgt dat

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A).$$

En op dezelfde wijze volgt dat $P(B|A) = P(B)$. Dus A en B zijn onafhankelijk.

Definitie De n gebeurtenissen $A_1, \dots, A_n$ zijn *onderling onafhankelijk* (*mutually independent*) als voor elke deelverzameling $\{A_{i_1}, A_{i_2}, \dots, A_{i_m}\}$ van $\{A_1, \dots, A_n\}$ geldt:

$$P(A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_m}) = P(A_{i_1})P(A_{i_2}) \dots P(A_{i_m}).$$

Voorbeeld Iemand beweert dat miljonair zijn onafhankelijk is van stemgedrag (in termen van links/rechts stemmen).

Er wordt bekend dat 56% van de mensen op een rechtse partij stemt, dat 2% miljonair is, en dat 1% van de mensen miljonair is en rechts stemt. Is de uitspraak waar?

$$0.56 \times 0.02 = 0.0112 \neq 0.01.$$

Ja, de uitspraak is waar. De twee kenmerken zijn niet onafhankelijk.

Voorbeeld Stel dat er 3 karakteristieken (zoals blauwe ogen of bloedtype C) zijn die wel of niet in een DNA monster kunnen voorkomen. A_i is de gebeurtenis dat karakteristiek i voorkomt in het monster. Vastgesteld is dat het percentage mensen met karakteristiek A_1, A_2, A_3 respectievelijk 10%, 20% en 5% is.

Aannemende dat de karakteristieken onafhankelijk zijn is de kans dat in een DNA monster van een persoon alle drie voorkomen $0.1 \times 0.2 \times 0.05 = 0.001$.

Bij het vaststellen van zekere kenmerken in het DNA van een verdachte wordt vaak onder de aanname van onafhankelijkheid gewerkt.

Dit is controversieel, want het is vaak onbekend of kenmerken afhankelijk zijn.

Stel bijvoorbeeld dat A_2 en A_3 wel afhankelijk zijn, maar onafhankelijk van A_1 , en dat de kans op A_3 gegeven A_2 0.6 is. Er geldt dat

$$P(A_2 \cap A_3) = P(A_3 | A_2)P(A_2) = 0.6 \times 0.2 = 0.12.$$

Met deze gegevens is de kans dat de karakteristieken alle drie voorkomen $0.1 \times 0.12 = 0.012$. Veel groter.

Stelling van Bayes

Stelling van Bayes



Thomas Bayes (1702-1761)

An Essay Towards Solving a Problem in the Doctrine of Chances

Stelling van Bayes

Merk op:

Uit

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad P(B|A) = \frac{P(A \cap B)}{P(A)}$$

volgt dat

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A).$$

Dat geeft

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}.$$

De kans $P(A|B)$ is uitdrukbaar in de kans $P(B|A)$.

Stelling van Bayes

Voorbeeld De kans dat een zekere kerncentrale oververhit raakt (O) is $1 \cdot 10^{-7}$, de kans dat er een lek (L) ontstaat is $1 \cdot 10^{-8}$, maar als er eenmaal een lek is, is de kans groot dat de centrale oververhit raakt: 0.1.

Wat is de kans dat de centrale gaat lekken als hij oververhit raakt?

$$P(L|O) = \frac{P(L \cap O)}{P(O)} = \frac{P(O|L)P(L)}{P(O)}.$$

Dus

$$P(L|O) = \frac{0.1 \cdot 1 \cdot 10^{-8}}{1 \cdot 10^{-7}} = \frac{1 \cdot 10^{-9}}{1 \cdot 10^{-7}} = 0.01.$$

Stelling van Bayes

Stelling (Speciaal geval van de Stelling van Bayes)

Als H en E twee gebeurtenissen zijn, dan geldt:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E|H)P(H) + P(E|\bar{H})P(\bar{H})}$$

$$P(\bar{H}|E) = \frac{P(E|\bar{H})P(\bar{H})}{P(E|H)P(H) + P(E|\bar{H})P(\bar{H})}.$$

Bewijs

Omdat

$$P(H|E) = \frac{P(H \cap E)}{P(E)},$$

volgt de stelling uit de volgende twee observaties:

$$P(E) = P(E \cap H) + P(E \cap \bar{H}) = P(E|H)P(H) + P(E|\bar{H})P(\bar{H}).$$

$$P(H \cap E) = P(E|H)P(H).$$

Het bewijs voor $\bar{H}$ is analoog.

Merk op: $P(H|E) + P(\bar{H}|E) = 1$.

Voorbeeld De kans dat een boek gekocht op internet illegaal is, is 0.00001.

De kans dat een boek dat gedrukt wordt bij een uitgeverij een bladzijde mist is 0.0001. De kans dat een illegale kopie een bladzijde mist is 0.02.

Wat is de kans dat een gekocht boek dat een bladzijde mist een illegale kopie is?

H : het boek is illegaal, $\bar{H}$: het boek is legaal. E : het boek mist een bladzijde.

De kans dat een boek dat een bladzijde mist een illegale kopie is:

$$P(H | E) = \frac{P(E | H)P(H)}{P(E | H)P(H) + P(E | \bar{H})P(\bar{H})} =$$
$$\frac{0.02 \times 0.00001}{0.02 \times 0.00001 + 0.0001 \times 0.99999} = 0.002.$$

Voorbeeld De kans dat een email spam is, is 99%. Een spamfilter is 98% betrouwbaar: 98% van alle spam wordt als spam geïdentificeerd en 2% van niet-spam wordt als spam geïdentificeerd.

Een email wordt als spam geïdentificeerd. Wat is de kans dat het spam is?

H : de email is spam. $\bar{H}$: de email is geen spam. E : de email is als spam geïdentificeerd.

$$P(H | E) = \frac{P(E | H)P(H)}{P(E | H)P(H) + P(E | \bar{H})P(\bar{H})} = \frac{0.98 \times 0.99}{0.98 \times 0.99 + 0.02 \times 0.01} = 0.9998.$$

De kans dat het spam is is met 0.0098 toegenomen door deze informatie, een toename van ongeveer 1%.

Als het filter slechts 49% betrouwbaar is neemt de kans dat een als spam geïdentificeerde email spam is zelfs af:

$$P(H | E) = \frac{0.49 \times 0.99}{0.49 \times 0.99 + 0.51 \times 0.01} = 0.9896.$$

Een test op *HIV* is 90% betrouwbaar: als een persoon *HIV* heeft is de kans op een positieve uitslag 0.9, en als een persoon geen *HIV* heeft is de kans op een positieve uitslag 0.1. De kans op *HIV* is 0.05.

Een dame ondergaat de test en de uitslag is positief. Wat is de kans dat zij *HIV* heeft?

De gevraagde kans is $P(HIV|POS)$. In de notatie van de Stelling van Bayes wordt dat $H = HIV$ en $E = POS$.

Met de Stelling van Bayes:

$$P(HIV|POS) = \frac{P(POS|HIV)P(HIV)}{P(POS|HIV)P(HIV) + P(POS|\overline{HIV})P(\overline{HIV})}.$$

Daarmee

$$P(HIV|POS) = \frac{0.9 \times 0.05}{0.9 \times 0.05 + 0.1 \times 0.95} = 0.32.$$

Voorbeeld Uit genetisch onderzoek blijkt dat de werking van een medicijn 1% of 10% moet zijn. Het eerste percentage is waarschijnlijker: de kans dat 1% het correcte percentage is, is 0.55, en voor 10% is dat dus 0.45.

Er wordt een steekproef van 100 zieken genomen en de werking van het medicijn in de steekproef is 11%. Welke werking is op grond van de steekproef het waarschijnlijkste: 1% of 10%?

H : de werking is 1%. $\bar{H}$: de werking is 10%, E : de werking in de steekproef is 11%.

$$P(H | E) = \frac{P(E | H)P(H)}{P(E | H)P(H) + P(E | \bar{H})P(\bar{H})} = \frac{\binom{100}{11}(0.01)^{11}(0.99)^{89} \times 0.55}{\binom{100}{11}(0.01)^{11}(0.99)^{89} \times 0.55 + \binom{100}{11}(0.1)^{11}(0.9)^{89} \times 0.45} = 6 \times 10^{-8}.$$

Dus $P(\bar{H} | E) = (1 - 6 \times 10^{-8}) > 6 \times 10^{-8}$, en de hypothese dat 10% het correcte percentage is, is dus het meest waarschijnlijk.

Voorbeeld Hetzelfde voorbeeld als op de vorige zijde, behalve dat de werking van het medicijn in de steekproef nu 6% is en de kans dat 1% het correcte percentage is, 0.995 is. Voor 10% is dat dus 0.005.

Welk percentage is nu het waarschijnlijkste?

H : de werking is 1%. $\bar{H}$: de werking is 10%, E : de werking in de steekproef is 6%.

$$P(H | E) = \frac{P(E | H)P(H)}{P(E | H)P(H) + P(E | \bar{H})P(\bar{H})} = \frac{\binom{100}{6}(0.01)^6(0.99)^{94} \cdot 0.995}{\binom{100}{6}(0.01)^6(0.99)^{94} \cdot 0.995 + \binom{100}{6}(0.1)^6(0.9)^{94} \cdot 0.005} = 0.6.$$

Dus $P(\bar{H} | E) = 0.4$, en de hypothese dat 1% het correcte percentage is, is het meest waarschijnlijk.

Hoewel de uitkomst van de steekproef meer in overeenstemming is met 10%, doet het feit dat die hypothese veel onwaarschijnlijker wordt geacht dat effect weer teniet.

Stelling van Bayes: gelijke kansen

Stelling Als H en E twee gebeurtenissen zijn en de kans op H en op $\bar{H}$ zijn gelijk, dan geldt:

$$P(H | E) = \frac{P(E | H)}{P(E | H) + P(E | \bar{H})}$$

$$P(\bar{H} | E) = \frac{P(E | \bar{H})}{P(E | H) + P(E | \bar{H})}.$$

Bewijs Het speciale geval van de Stelling van Bayes heeft de vorm:

$$P(H | E) = \frac{P(E | H)P(H)}{P(E | H)P(H) + P(E | \bar{H})P(\bar{H})}$$

$$P(\bar{H} | E) = \frac{P(E | \bar{H})P(\bar{H})}{P(E | H)P(H) + P(E | \bar{H})P(\bar{H})}.$$

Als $P(H) = P(\bar{H})$ kunnen de factoren $P(H)$ en $P(\bar{H})$ tegen elkaar weggestreept worden.

Finis