Data-Science 1

kansen

Karel de Grote Hogeschool

06/03/2024

Oefeningen huiswerk

- heeft het zin om een regressielijn te bepalen tussen schoenmaat en lengte?
- Kijk eens naar de correlatie tussen opwarming en zakgeld. Wat betekent dat? Heeft het zin om hier een regressielijn te bepalen?

Oefeningen huiswerk

- bereken een lineaire regressie die de schoenmaat voorspelt adh van de lengte
  - verwijder eerst de uitschieters
  - wat is de vergelijking van de rechte?
- welke schoenmaat voorspel je voor iemand van 180cm groot?
- wat is de gemiddelde fout op de voorspelling? Is dit veel of weinig?
- wat is de verklaarde variantie? Wat betekent dit?
- doe eens een logaritmische regressie. Welke vergelijking kom je uit? Geeft dit een beter of slechter resultaat dan een lineaire?

Oefeningen huiswerk

- In het leerboek “Edexcel GCSE Mathematics” van Keith Pledger wordt beweert dat (voor jongens) de lengte in cm gelijk is aan 5,3 maal (Engelse) schoenmaat plus 133. Welke formule bekom je voor de informatica studenten?

  je kan EU maten omzetten naar UK met:

  UK = EU - 34

Inhoud

- welk probleem willen we oplossen?
- kansen
- rekenen met kansen
- regel van Bayes

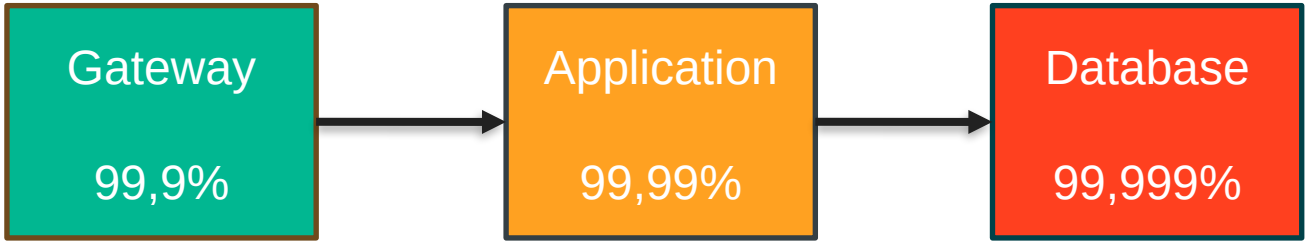
Welk probleem?

High availability server

- de beschikbaarheid van een server
  - reliability: de **kans** dat de server niet faalt
  - maintainability: de **kans** dat de server na falen succesvol hersteld wordt
  - availability: de **kans** dat een server op het gegeven moment niet faalt en niet wordt hersteld na een faling

High availability server

- de availability van een IT service hangt af van de availability van elk van de componenten
- Voorbeeld – Kubernetes cluster:

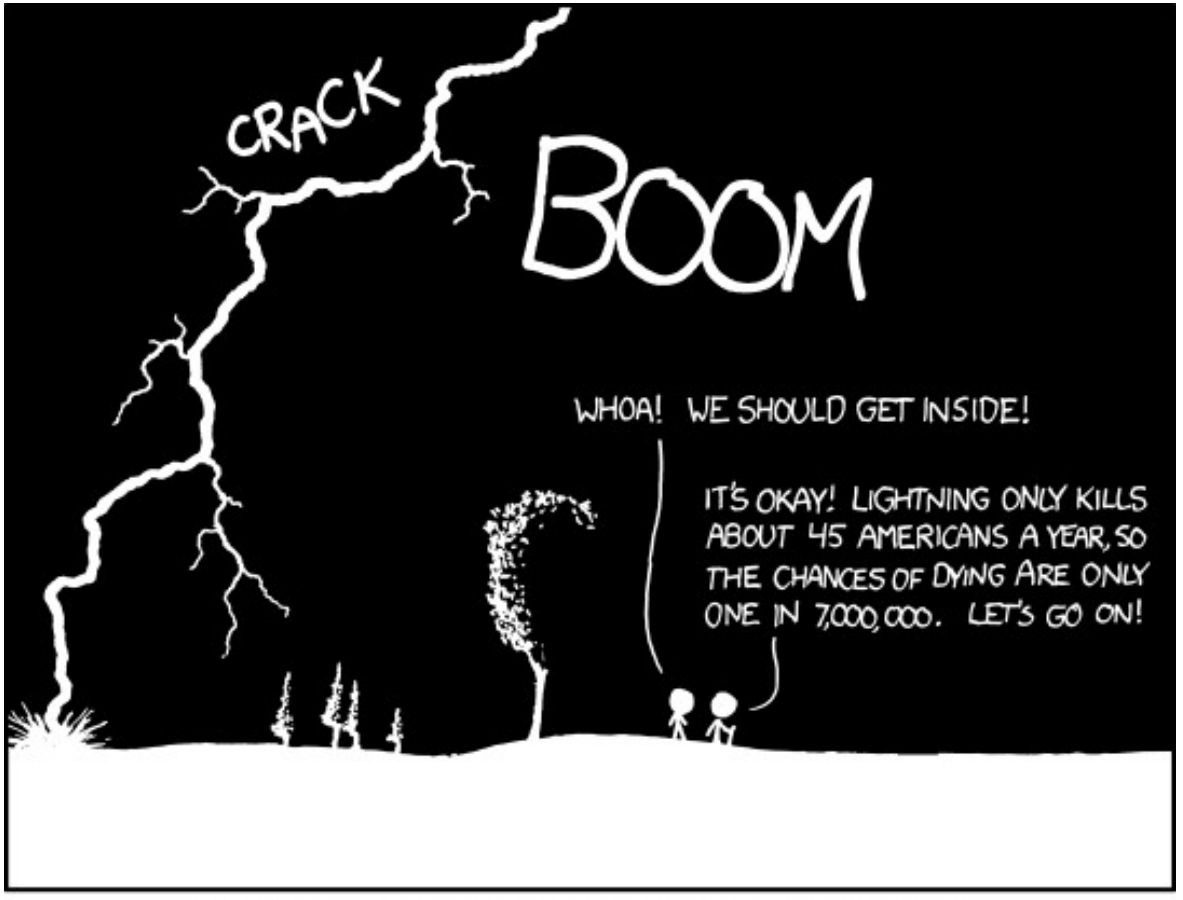


- availability cluster = 0.999 x 0.9999 x 0.99999 = 0.9989
- 99.89 % = 10 uur downtime / jaar

Kansen

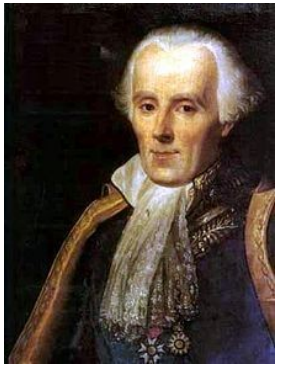
Kansen zijn raar...

- slaagkans
  - 2 mogelijkheden, dus ...
  - vorig jaar waren 70 van de 210 geslaagd, dus ...
  - voorbije 10 jaar was het slaagpercentage gemiddeld 25%, dus ...
  - als je KSO, ASO, TSO, BSO gevolgd hebt, is je slaagkans dan anders?
  - op het einde van het jaar is je slaagkans 100% of 0% (naargelang je resultaat)
- in dit hoofdstuk: enkel kansen in omstandigheden die duidelijk afgelijnd zijn



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

Kansen volgens Laplace



- "afgelijnd experiment"
- gegeven: verzameling mogelijke uitkomsten en verzameling "gunstige" uitkomsten
- gevraagd: wat is de kans op een gunstige uitkomst?
- voorbeeld: zak knikkers met 20 blauwe en 30 rode. Ik kies een knikker. Wat is de kans op een rode knikker?

Kansen volgens Laplace

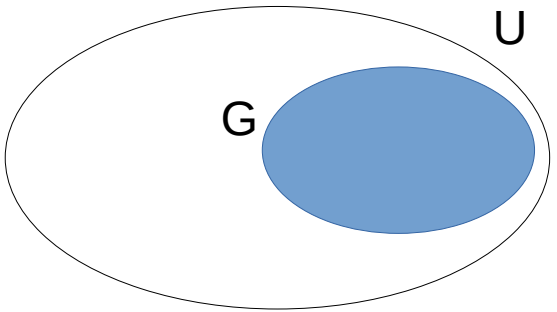
- P(gebeurtenis) = aantal gunstige uitkomsten / aantal mogelijke uitkomsten
  - notatie P() voorbeeld: P(regen), P(ziek), ...
  - 0 ≤ kans ≤ 1
  - kans * 100 = kans in percent (zet er dan altijd % achter!)

Kansen volgens Laplace

- voorbeeld:
  - aantal studenten TI vorig jaar = 210
  - aantal geslaagde studenten TI vorig jaar = 70
  - wat is de kans dat een willekeurige student van vorig jaar geslaagd was?
  - wat is de kans dat een willekeurige student dit jaar geslaagd zal zijn?

Kansen volgens Laplace

- met verzamelingen

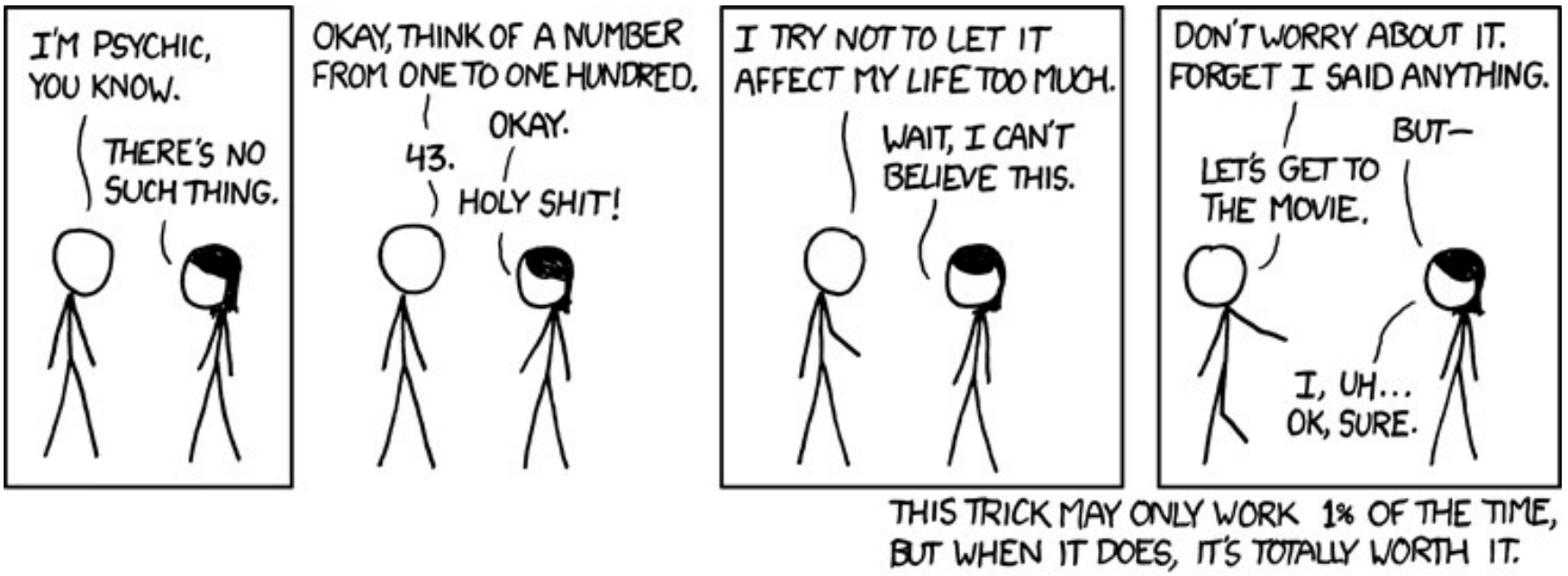


- P(G) = #G / # U

Voorbeeld

- twee dobbelstenen: wat is de kans om 7 te gooien?

	Gebeurtenis A	#A
2	(1,1)	1
3	(1,2); (2,1)	2
4	(1,3); (2,2); (3,1)	3
5	(1,4); (2,3), (3,2); (4,1)	4
6	(1,5); (2,4); (3,3); (4,2); (5,1)	5
7	(1,6); (2,5); (3,4); (4,3); (5,2); (6,1)	6
8	(2,6); (3,5); (4,4); (5,3); (6,2)	5
9	(3,6); (4,5); (5,4); (6,3)	4
10	(4,6); (5,5); (6,4)	3
11	(5,6); (6,5)	2
12	(6,6)	1
	#U (= TOTAAL)	**36**


I'M PSYCHIC, YOU KNOW.
THERE'S NO SUCH THING.
OKAY, THINK OF A NUMBER FROM ONE TO ONE HUNDRED.
43.
OKAY.
HOLY SHIT!
I TRY NOT TO LET IT AFFECT MY LIFE TOO MUCH.
WAIT, I CAN'T BELIEVE THIS.
DON'T WORRY ABOUT IT. FORGET I SAID ANYTHING.
LET'S GET TO THE MOVIE.
BUT—
I, UH... OK, SURE.
THIS TRICK MAY ONLY WORK 1% OF THE TIME, BUT WHEN IT DOES, IT'S TOTALLY WORTH IT.

Kansen als relatieve frequentie

- gegeven: zak met 20 blauwe en 30 rode knikkers
- gevraagd: als ik 1000 keer een knikker neem (en terug leg), wat is dan de verwachte relatieve frequentie?

Kans als relatieve frequentie

Steekproef: 10, 100 en 1000 keer:

steekproef lengte	rel. freq. blauw	rel. freq. rood
10	0,7	0,4
100	0,44	0,56
1000	0,392	0,608
10000	0,4019	0,5981

Kans is dus de relatieve frequentie van een (theoretische) oneindige steekproef

Voorwaardelijke kansen

- kansen kunnen afhangen van bepaalde kennis
- voorbeeld: wat is de kans dat een smartphone nog werkt als ik weet dat die in het water is gevallen?
- notatie: P(werkt | water)
- als P(werkt) = P(werkt | water) dan zeggen we dat de gebeurtenissen “onafhankelijk” zijn

Kans als relatieve frequentie

Je kan voorwaardelijke kansen soms ook gemakkelijk uit een kruistabel aflezen:

	Wit merk (White label)	Geen wit merk (Private label)	
Slechte koeling	1498	1513	**3011**
Goede koeling	504	6485	**6989**
	2002	**7998**	**10000**

Wat is de kans dat een PC van een wit merk uit deze steekproef een slechte koeling heeft?

Rekenen met kansen

2

De inverse

- P(niet G) = 1 - P(G)
- bv: kans dat het regent is 55%, wat is de kans dat het niet regent?

  P(regent niet) = 1 - P(regent) = 1 - 0,55

De productregel

- voorbeeld: groep kinderen, 40% meisjes, 60% jongens, 10% draagt een bril
- dus:
  - P(jongen)=
  - P(meisje)=
  - P(bril)=
- gevraagd: kans op een brildragend meisje
  - P(meisje EN bril) = ?

De productregel

- als er evenveel meisjes als jongens een bril dragen:

  P(meisje EN bril) = P(meisje) * P(bril)

- maar... wat als alle brildragers jongens zijn?

De productregel

- meer algemene formule:

  P(meisje EN bril) = P(meisje) * P(bril | meisje)

- wat is P(bril | meisje) ???
- stel dus dat 20% van de meisjes een bril draagt, dan ...
- opmerking: je kan het ook zo vinden (afhankelijk van de gegevens die je hebt):

  P(meisje EN bril) = P(bril) * P(meisje | bril)

De productregel

- met kruistabel: stel dat er 100 kinderen zijn

	jongens	meisjes	totaal
bril			10
geen bril			
totaal	60	40	100

20% van de meisjes draagt een bril...

De productregel

- met kruistabel: stel dat er 100 kinderen zijn

	jongens	meisjes	totaal
bril	2	8	10
geen bril	58	32	90
totaal	60	40	100

20% van de meisjes draagt een bril...

De productregel

- gebeurtenissen zijn "onafhankelijk" als P(A|B)=P(A)
- voorbeeld: kans dat iemand een rode T-shirt draagt en kans dat iemand goed is in wiskunde
- P(rood EN wiskunde)
  = P(rood) * P(wiskunde | rood)
  = P(rood) * P(wiskunde)

De somregel

- voorbeeld: wat is de kans dat een willekeurig kind een meisje is OF een bril draagt (of allebei)?

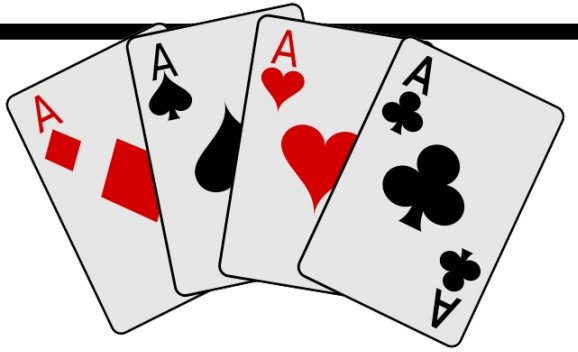
	jongens	meisjes	totaal
bril	2	8	10
geen bril	58	32	90
totaal	60	40	100

De somregel

- P(meisje OF bril)
  = P(meisje) + P(bril) - P(meisje EN bril)
  = 0,4 + 0,1 - 0,08 = 0,42 = 42%
- algemeen:

  P(A of B) = P(A) + P(B) - P(A en B)

Voorbeeld kaarten



- gegeven: boek kaarten, trek er 1 uit
- wat is de kans dat die kaart een aas is en een harten?

  P(aas en harten)
  = P(aas) * P(harten | aas)
  = 4/52 * 1/4 = 1/52

- wat is de kans dat die kaart een aas is of een harten?

  P(aas of harten)
  = P(aas) + P(harten) - P(aas en harten)
  = 4/52 + 13/52 - 1/52 = 16/52

Voorbeeld gaming

- 3 spelers proberen vijand uit te schakelen
  - speler1: 1/2 kans om te raken
  - speler2: 1/3 kans om te raken
  - speler3: 1/4 kans om te raken

- wat is de kans dat de vijand overleeft?

Voorbeeld gaming

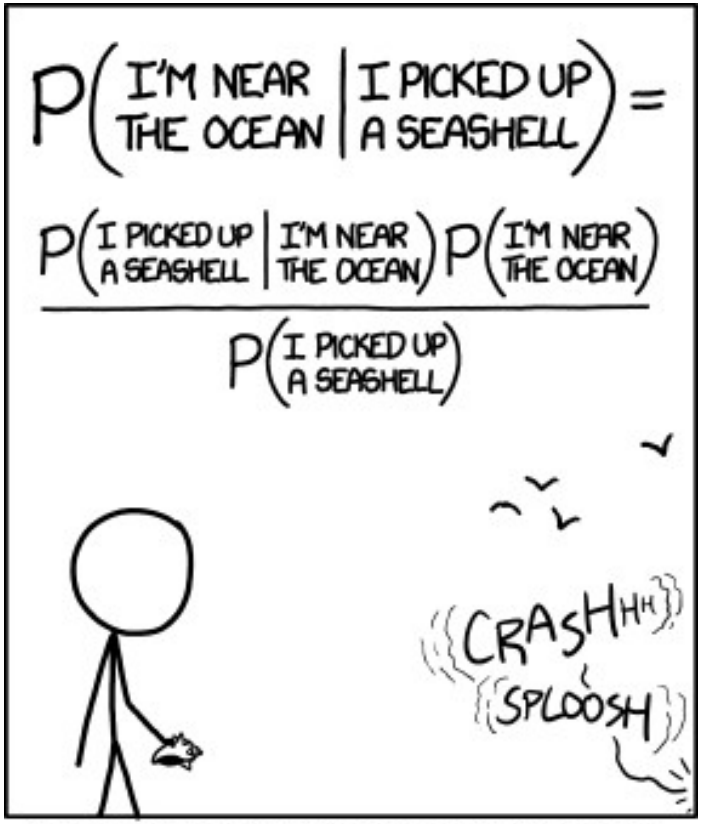
- kans op overleven = speler 1 mist EN speler 2 mist EN speler 3 mist
- spelers schieten 1 maal, onafhankelijk van elkaar
- P(overleven) =
  P(speler1 mist) *
  P(speler2 mist) *
  P(speler3 mist)
- P(overleven) = 1/2 * 2/3 * 3/4 = 0,25
- dus...

Regel van Bayes

De regel van Bayes

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

- voorbeeld:
  - kans dat student slaagt = 80%
  - kans dat student zich bezat als hij gebuisd is = 90%
  - kans dat student zich bezat = 60%
- we zien een zatte student: wat is de kans dat hij gebuisd is?



STATISTICALLY SPEAKING, IF YOU PICK UP A SEASHELL AND *DON'T* HOLD IT TO YOUR EAR, YOU CAN PROBABLY HEAR THE OCEAN.

Machine learning met Bayes

- vooral interessant bij analyseren van tekst
  - is een bepaalde tekst verdacht of niet?
  - is een bepaalde tekst spam of niet?
  - bevat een tekst een positieve of negatieve emotie?
  - ...

Voorbeeld: spam

- gegeven een tekst, wat is de kans dat deze spam is?

  P(spam | woorden)

  = P(woorden | spam) * P(spam) / P(woorden)

- gegeven: teksten uit het verleden met aanduiding of ze spam waren of niet

Voorbeeld spam

- we benaderen de kansen met relatieve frequenties (“waarschijnlijkheid” of “likelyhood”):
  - P(spam) ~ #spam / #mails
  - P(woorden) ~ P(woord1) * P(woord2) * ... * P(woordn)
  - P(woorden | spam) ~ P(woord1 | spam) * P(woord2 | spam) * ... * P(woordn | spam)

4

Samenvatting

Samenvatting

- P(niet A) = 1 - P(A)
- P(A en B) = P(A) * P(B|A)
- P(A of B) = P(A) + P(B) - P(A en B)
- P(A|B) = P(B|A) * P(A) / P(B)

Oefeningen

Oefeningen

- Klassikaal
  - 3 politici hebben resp. 30%, 20% en 50% kans dat ze verkozen worden (V1, V2, V3)
  - de kans dat de politici de belastingen verlagen is resp. 50%, 40% en 30%
  - wat is de kans dat de belastingen verlaagd worden?

Oefeningen

$$
\begin{aligned}
P\left(B_{verlaagd}\right) \quad &= P\left(V_1\right)\cdot P\left(B_{verlaagd}\middle|V_1\right) \\
&+ P\left(V_2\right)\cdot P\left(B_{verlaagd}\middle|V_2\right) \\
&+ P\left(V_3\right)\cdot P\left(B_{verlaagd}\middle|V_3\right)
\end{aligned}
$$

Oefeningen

- Klassikaal
  - na de verkiezingen worden de belastingen verlaagd
  - hoeveel kans is er dat dat komt doordat politicus 3 verkozen werd?

Oefeningen

$$P(V_3|B_{verlaagd}) = \frac{P(V_3) \cdot P(B_{verlaagd}|V_3)}{P(B_{verlaagd})}$$

Oefeningen

- Zie Canvas