Casper Albers
Psychometrie & Statistiek
Rijksuniversiteit Groningen
c.j.albers@rug.nl

Column Casper grijpt een kans

Er is er een jarig

Caspers Albers zal vanaf nu op regelmatige basis in dit blad een column schrijven over alledaagse statistische onderwerpen.

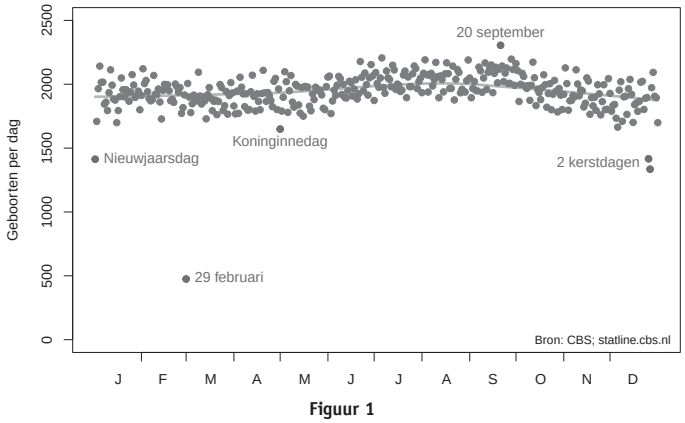
Dit wordt mijn eerste bijdrage aan NAW en het lijkt me gepast dat ik die dan maar besteedt aan een van de bekendere problemen uit de statistiek en kansrekening: de verjaardagenparadox. *Hoe groot moet een groep willekeurige personen zijn zodat de kans dat er minstens twee personen op dezelfde dag jarig zijn minimaal 50 procent is?* Voor het beantwoorden van deze vraag wordt doorgaans de aanname gemaakt dat er elke dag precies evenveel mensen jarig zijn (en dat 29 februari niet bestaat).

Vervolgens is het beantwoorden van de vraag een kwestie van elementaire kansrekening (althans, elementair voor het lezerspubliek van dit blad). De kans dat in groep van één persoon iedereen een unieke verjaardag heeft, is vanzelfsprekend 1. De kans dat een nieuwe persoon die wordt toegevoegd aan de groep ook een unieke verjaardag heeft, is $364/365$; bij een derde persoon is dit $363/365$, et cetera. Voor een groep van k personen, is de kans op k unieke verjaardagen dan $365!/(365^k \times (365-k+1)!)$. Bij $k=23$ is deze kans kleiner dan 50 procent en de kans op overlappende verjaardagen dus groter dan 50 procent ($50{,}7$ procent om wat preciezer te zijn).

Je kan hier vervolgens hele interessante wiskunde mee doen — de Taylorontwikkeling is bijvoorbeeld verrassend accuraat — maar dat ga ik niet doen: ik wil me buigen over de statistiek rond dit probleem: die aanname, dat elke dag van het jaar evenveel jarigen kent, klopt die wel? En als die aanname niet klopt — op de ene dag zijn meer jarigen dan op de andere — wat heeft dat dan voor gevolgen voor de groepsgrootte die nodig is om minimaal 50 procent kans op een dubbele verjaardag te hebben? Of je juist grotere of kleinere groepen kan verwachten is intuïtief goed te beredeneren door aan een extreem scenario te denken: als iedereen op dezelfde dag jarig is (dus maximale niet-uniformiteit), dan is $k=2$ al voldoende. De groepsgrootte zal dus kleiner worden (of, doordat we met gehele getallen werken, gelijk blijven).

Het CBS houdt minutieus bij hoeveel geboortes er per dag zijn, het is bijna alsof er in elke kraamkamer een statisticus staat te turven. In de maanden juli tot en met oktober ligt het geboortecijfer per dag in Nederland bijna 10 procent boven dat van de andere maanden. En dan zijn er ook nog speciale dagen (zoals 1 januari en 25 december) waarop er wel natuurlijke bevallingen zijn (baby's houden geen rekening met feestdagen), maar geen bevallingen door geplande keizersneden.

Ik heb de CBS-gegevens van de jaren 2010–2013 bestudeerd (een vierjarige periode zodat er een jaar met, en drie jaren zonder 29 februari in zitten). Het aantal geboortes per dag varieert van 302 (op 15 december 2013) tot 625 (22 september 2010): niet bepaald uniform. Over deze vier jaren heen blijkt 29 februari vanzelfsprekend de minst populaire dag om te bevallen; de vruchtbaarste dagen zijn 20 september en 5 juli (geen idee waarom), zie Figuur 1. Gemiddeld worden er zo'n $485{,}7$ baby's per dag geboren, met een standaarddeviatie van $36{,}1$. Als we nu de aanname maken dat deze vier jaar een goede afspiegeling vormen van de verdeling van geboortes over de jaren, kunnen we een realistischere berekening maken. (Al zal de echte verjaardagsverdeling ongetwij-



Figuur 1

feld niet exact hetzelfde zijn als de verdeling van geboortes over 2010–2013.)

Je zou dit probleem — in theorie — wiskundig aan kunnen pakken. De kans op een dubbele verjaardag is één min de kans op enkel unieke verjaardagen. Voor elk van de 366 dagen is er nu een aparte kans, $p_1,\ldots,p_{366}$ (met restrictie $\sum_i p_i = 1$), op een verjaardag. Voor een groep van k personen, zijn er $(366-k+1)!$ verschillende combinaties van verjaardagen zodat er geen dubbele vergaderingen zijn — een gigantisch groot aantal. Voor elk van die combinaties is uit te rekenen wat de kans erop is via de multinomiale verdeling, in dit geval:

$$\mathrm{P}(x_1,\ldots,x_k) = k! \prod_{i=1}^{k} p_{x_i},$$

en omdat de p's gemiddeld $1/366$ zijn, levert dit een microscopisch klein getal op bij grotere k. Uiteindelijk moet je een gigantisch aantal microscopische getallen optellen. Theoretisch kan het, maar praktisch is het niet.

Een veel praktischere oplossing is om deze kans te benaderen met brute rekenkracht wel: met een paar regels programmeercode heb je binnen enkele seconden een miljoen groepjes van 23 personen gesimuleerd en geturfd in hoeveel groepjes personen een verjaardag delen. Die proportie gedeelde verjaardagen levert een goede schatting op van de gezochte kans, alsmede de foutenmarge die je door het toeval krijgt. (Deze is bij een miljoen replicaties al erg klein, maar als je niet snel tevreden bent kan je gewoon nog even doorrekenen. Bij elke verviervoudiging van het aantal rekenstappen, halveert de foutenmarge.) Binnen een minuut heb je het ook voor groepen van $22, 21, \ldots, 1$ persoon. En wat blijkt uit deze simulaties: de groepsgrootte k is nog steeds 23. Bij $k = 22$ is de kans op gedeelde verjaardagen $47{,}6$ procent, bij $k = 23$ is dit $50{,}8$ procent. Het verschil met $50{,}7$ procent dat je krijgt met de totaal onrealistische aanname van gelijke verdelingen van verjaardagen is dus verwaarloosbaar klein.

Dit is een interessante eigenschap die vaak opgaat in de statistiek: om je model op te stellen moet je aannames maken die vrij onrealistisch ogen. Vervolgens kan je met dat model de zaken die je wilt weten uitrekenen, maar het blijft knagen omdat je die aanname moest maken. Voor veel statistische modellen geldt echter dat ze erg robuust zijn tegen schendingen van de aanname: als je het niet te bont maakt (dus niet iedereen op 1 januari geboren laten zijn), krijg je prima antwoorden, ook als je aanname niet helemaal klopt. En dat is wel zo fijn: want dan kan je statistiek ook gebruiken in de situaties waarin onrealistische aannames niet opgaan.