Het leidende-cijfer-fenomeen

Simon van der Salm

Over de logaritmische distributie van de eerste cijfers van getallen

1. De Wet van Benford oftewel de First Digit Law

In 1881, toen logaritme-tabellen nog duur waren en vaak door studenten uit een bibliotheek werden geleend, deed de Amerikaanse astronoom en wiskundige *Simon Newcomb* een merkwaardige ontdekking. Hij merkte op dat de eerste bladzijden van boeken met logaritme-tabellen veel vuiler waren dan de bladzijden aan het einde.
De verzamelaars van (liefst oude en veel gebruikte) logaritme-tabellen onder de lezers van dit artikel raad ik aan hun collectie te doorzoeken. Wellicht vinden ze enig bewijs voor de curieuze waarneming van Newcomb.
Newcomb trok schertsend de conclusie dat veel van zijn studenten kennelijk werden aangetrokken door de logaritme-tabellen, dat veel van hen zeer intensief de eerste pagina's hadden bestudeerd en dat blijkbaar veel studenten na eerste lezing er de brui aan hadden gegeven, net alsof logaritmetabellen tot die categorie goedkope romannetjes behoorden, waarvan na lezing van de eerste bladzijden al duidelijk is dat verder lezen verspilling van kostbare tijd betekent.
Newcomb bedacht dat natuurwetenschappers en ingenieurs kennelijk een grotere behoefte hebben aan getallen die met een 1 of een 2 beginnen, dan aan getallen die met een 8 of een 9 beginnen. Maar dat betekent dat getallen die met een laag cijfer beginnen vaker moeten voorkomen dan getallen met een hoog cijfer! En inderdaad, getallen hebben veel vaker 1 als leidend cijfer dan 2, vaker 2 dan 3, enzovoorts. Of wat deftiger gezegd: de distributie van het leidende cijfer van getallen is niet uniform.
Newcomb vond zelfs de formule die, voor het decimale stelsel, de kans geeft dat een willekeurig gekozen getal het leidende cijfer d (= 1, 2, ..., 9) bezit:

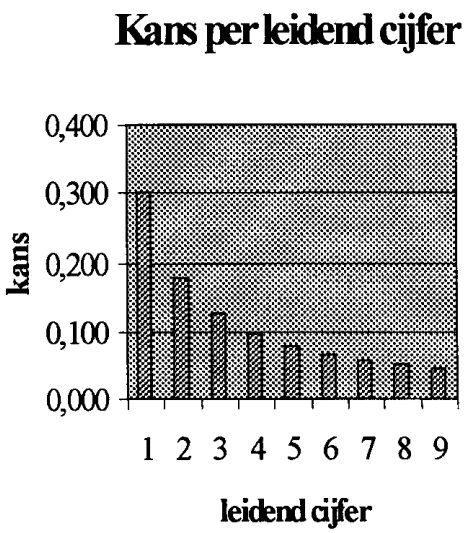
$$P_{(\text{leidende cijfer} = d)} = {}^{10}\log(1 + 1/d) \qquad (1.1)$$

(P staat voor *probabiliteit*, dus voor kans).

De volgende tabel geeft de kans weer dat een specifiek cijfer het leidende cijfer van een getal zal zijn.

Cijfer	Kans
1	0,301
2	0,176
3	0,125
4	0,097
5	0,079
6	0,067
7	0,058
8	0,051
9	0,046

In figuur 1.1 staat de kans per leidend cijfer uitgezet in een grafiek



Figuur 1.1

De kans dat een getal met een 1 begint, is dus - volgens deze formule - ongeveer 0,301; de kans dat een getal met een 9 begint is ongeveer 0,046. De kans dat een getal met een 1 begint is dus veel groter dan de kans dat een getal met een 9 begint.
Men kan dit ook zo interpreteren: als je beschikt over een grote verzameling ‘willekeurig’ gekozen getallen, dan begint circa 30 procent van de getallen met een 1; 17,6 procent met een 2, enzovoort.

Newcomb gaf echter geen statistisch bewijs voor de juistheid van zijn formule. Bovendien raakte de formule in vergetelheid. Tot 1938, toen *Frank Benford* dezelfde ontdekking als Newcomb deed en dezelfde formule vond voor de distributie van het leidende cijfer van getallen. Sindsdien wordt gesproken over het *fenomeen van Benford* en over de *wet van Benford.* De wet van Benford wordt ook wel met de *First Digit Law* aangeduid. Bovenstaande formule 1.1 wordt meestal de formule van Benford genoemd, maar het is eigenlijk aardiger en juister hem de *formule van Newcomb-Benford* te noemen.

Benford baseerde zijn formule op het onderzoek van een enorme verzameling numerieke gegevens, zoals de oppervlakte van honderden rivieren, de soortelijke warmte van duizenden chemische componenten, tabellen met wortels uit getallen, duizenden getallen die in kranten en tijdschriften waren gepubliceerd, enzovoorts. En inderdaad leiden de gegevens die Benford verzamelde tot de overtuigende (empirisch gevonden) conclusie dat de wet van Benford moet gelden.

2. De First Digit Law voor gebruikers van rekenlinialen

Het belang van de wet van Benford voor gebruikers van logaritme-tabellen is evident: u kunt beter de laatste bladzijden van uw logaritmetabellen kwijtraken dan de eerste bladzijden. (Ervan uitgaand dat uw tabellen met de logaritme van 100 beginnen en met die van 999 eindigen)

De formule van Newcomb-Benford voor de kans dat een getal met een specifiek cijfer d begint komt overeen met de formule voor de afstand tussen opeenvolgende natuurlijke getallen op een logaritmische schaal. Zie bijvoorbeeld de C- en D-schalen die op iedere gewone rekenliniaal voorkomen. Op de C-schaal komen de natuurlijke getallen 1 tot en met 10 voor, maar de afstand tussen twee van zulke opeenvolgende getallen wordt kleiner naarmate die getallen groter zijn.

Het getal 1 op de C-schaal staat bij het punt 0 van een achterliggende lineaire schaal. Het getal 10 staat bij het getal 1 van de lineaire schaal. (Deze lineaire schaal is de schaal L die op de meeste linialen voorkomt).
Is X een getal op de C-schaal, dan is de werkelijke afstand tot het nulpunt van de lineaire schaal gelijk aan de logaritme van X.
De afstand tussen 1 en 2 op de C-schaal, gemeten langs de lineaire schaal, is:

$$\log 2 - \log 1 = \log 2/1$$
$$= \log(1 + 1/1) = \log 2 \;.\; 0{,}301$$

De afstand tussen 2 en 3 op de C-schaal, gemeten langs de lineaire schaal, is:

$$\log 3 - \log 2 = \log 3/2 = \log(1 + 1/2) \;.\; 0{,}176$$

Enzovoorts.
Uiteindelijk vinden we voor de afstand tussen 9 en 10 op de C-schaal:

$$\log 10 - \log 9 = \log(1 + 1/9) \;.\; 0{,}046$$

We vinden dus hetzelfde rijtje getallen als in de eerdere tabel.

De getallen op de C-schaal waarmee de gebruiker van de rekenliniaal het meest rekent komen dus voor in het grootste stuk, links van de liniaal; de getallen waaraan hij de minste behoefte heeft, komen voor in het kleinste stuk, rechts van de rekenliniaal. We vinden immers in die stukken respectievelijk de getallen die met 1 beginnen en de getallen die met 9 beginnen. Verrassend is dat dus op de C-schaal (en tevens op enkele andere schalen) onbedoeld de meeste ruimte is gecreëerd voor de getallen die we het vaakst nodig hebben.

Voor getalstelsels met een ander grondtal dan 10 geldt een variant van de eerder vermelde formule van Newcomb-Benford (formule 1.l):

$$P_{(\text{leidende cijfer} = d)} = {}^{\$}\log(1 + 1/d) \qquad (2.1)$$

Hierin is:

$ het grondtal van het betreffende getalstelsel,

d JJn van de cijfers 1, 2, $-1.

3. Pogingen tot verklaring

Sinds het einde van de jaren dertig hebben talloze (amateur- en beroeps-)wiskundigen, statistici en ingenieurs getracht de empirische formule van Newcomb-Benford te verklaren.

Veel verklaringen zijn mathematisch ondeugdelijk en tegen veel andere verklaringen is veel in te brengen omdat ze gebaseerd zijn op hypotheses waarvoor geen bewijs wordt aangevoerd of kan worden. Zie het artikel van Raimi *[2] . (Zie referenties *[2] aan eind van dit artikel).

De meest populaire hypothese is het *schaal-invariantie-argument*, namelijk het axioma dat de distributie van het leidende cijfer van bijvoorbeeld de lengte van rivieren niet mag veranderen als we de meeteenheid veranderen van bijvoorbeeld een kilometer in een mijl.
Of een ander voorbeeld:
Als we een groot aantal temperaturen gemeten in graden Celsius bestuderen, dan zullen we opmerken dat ongeveer 30 procent van de getallen met 1 begint, dat ongeveer 17,6 procent met 2 begint, enzovoorts.

Rekenen we die temperaturen om naar graden Fahrenheit, dan moet nog steeds ongeveer 30 procent van de nieuwe getallen beginnen met 1, van die getallen moet 17,6 procent met 2 beginnen, enzovoorts. Deze hypothese ligt natuurlijk erg voor de hand, alleen het is niet zonder meer duidelijk op grond van welke *wiskundige* argumenten ze zou moeten gelden.

Een correcte wiskundige behandeling van het leidende-cijfer-probleem kan de geïnteresseerde lezer vinden in een artikel, gepubliceerd in 1995, van de Amerikaanse wiskundige *T P. Hill,* die verbonden is aan het Georgia Institute of Technology. (Zie referenties *[1] aan eind van dit artikel).

4. De General Significant-Digit Law
Hill geeft in zijn artikel *[1] bovendien de formule voor de *General Significant-Digit Law (GSDL).*

Een voorbeeld laat zien hoe deze wet luidt. We weten dat ð ongeveer 3,14 is. Als we de leidende cijfers van een willekeurig decimaal genoteerd getal met drie of meer cijfers bestuderen, hoe groot is dan de kans dat het betreffende getal met de cijfers 3, 1 en 4 begint?
De GSDL zegt:

$$P_{(\text{getal begint met } 314)} = {}^{10}\log (1 + 1/314) \;.\; 0{,}0014$$

De relatie tussen de GSDL en de C-schaal van de rekenliniaal is gemakkelijk te vinden als we ons in eerste instantie tot twee cijfers beperken, bijvoorbeeld tot de cijfers 1 en 2. De kans dat een getal met de cijfercombinatie 12 begint is:

$$^{10}\log(1 + 1/12) = {}^{10}\log (13/12)$$

$$= {}^{10}\log (1{,}3) - {}^{10}\log (1{,}2) \;.\; 0{,}035$$

Kennelijk is deze kans precies gelijk aan de werkelijke afstand tussen de getallen 1,2 en 1,3 op de C-schaal. (Dat wil dus zeggen: afstand gemeten door middel van de L-schaal).

Op overeenkomstige wijze kunnen we vinden dat de kans dat een willekeurig gekozen getal met de bovenstaande cijfercombinatie 314 begint gelijk is aan de werkelijke afstand tussen de getallen 3,14 en 3,15 op de C-schaal.

Eén van de merkwaardigste gevolgen van de GSDL is, dat de leidende cijfers in getallen *niet onafhankelijk* van elkaar zijn, in tegenstelling tot hetgeen men geneigd is te vermoeden. Maar het bestaan van de wet van Benford laat al zien dat de intuïtie in het geval van de distributie van cijfers niet erg betrouwbaar is.
Als een getal bijvoorbeeld met een 1 begint, dan is de *voorwaardelijke kans* dat het tweede cijfer een 2 is gelijk aan de kans dat het getal met 12 begint gedeeld door de kans dat het getal met een 1 begint, dus ongeveer 0,035/0,301 . 0,115.
Maar de *onvoorwaardelijke kans* dat het tweede cijfer 2 is, is hieraan ongelijk, namelijk ongeveer 0,109. Dit laatste getal kan men vinden door de kansen op de leidende cijfercombinaties 12, 22, 92 bij elkaar op te tellen.
Hieruit moeten we de volgende conclusie trekken: als een getal met een 1 begint is de kans dat het volgende cijfer een 2 is groter dan de kans dat het tweede cijfer een 2 is, ongeacht het eerste cijfer.

5. Falsificatie van numerieke gegevens
Een van de opmerkelijkste toepassingen van de bijzondere en de algemene wet van Benford ligt in het detecteren van fraude in een grote verzameling numerieke data, bijvoorbeeld de jaarcijfers en andere getallen die een bedrijf aan de belastingdienst verstrekt. Een elementaire, met een computer uitgevoerde statistische analyse kan tamelijk eenvoudig aantonen of, en in welke mate, de numerieke data voldoet aan de

bijzondere of algemene wet van Benford. Als de data niet min of meer aan die wetten voldoet, dan kan dat een aanwijzing zijn dat er met de data is geknoeid. Helaas: na enige tijd zullen vermoedelijk de in numerieke wiskunde geïnteresseerde fraudeurs ook op de hoogte zijn van de GSDL en juist deze GSDL gebruiken om hun fraude op intelligente wijze te verdoezelen.

Een andere belangrijke toepassing van de wet van Benford vinden we in de foutenanalyse van numerieke computerprogramma's die overdadig gebruik maken van zogenaamde floating-point berekeningen.

Referenties:

*[1] Hill, T.P., The Significant-digit Phenomenon,
American Mathematical Monthly 102, April 1995, 322-327

*[2] Raimi, R. A., The First Digit Phenomenon,
American Mathematical Monthly 83, Aug. 1976, 521-538
(Raimi geeft een uitgebreid en interessant overzicht van allerlei ingenieuze verklaringen voor de wet van Benford).