Wanneer je van een grote hoeveelheid getallen of data het eerste cijfer neemt, dan zou je kunnen denken dat je dan even vaak een 1, als een 2, als een 3 enzovoort tegenkomt. Maar net als bij de beginletters van woorden blijkt dat niet zo te zijn. **Simon van der Salm** legt uit hoe de begincijfers van getallen verdeeld zijn en wat je aan die kennis hebt.

Benfords logaritmische distributie van cijfers

De niet-uniforme verdeling van het leidende cijfer

Bij het bestuderen van grote verzamelingen data, van bijvoorbeeld fysische meetresultaten, valt al snel op dat er meer getallen met een 1 beginnen dan met een 2, dat er meer getallen met een 2 beginnen dan met een 3, enzovoorts. Kennelijk is de kans dat een getal met een 1 begint groter dan dat het getal met één van de andere cijfers begint. De kansverdeling van het leidende cijfer (het meest linkercijfer) van decimaal gerepresenteerde getallen is blijkbaar niet uniform. Maar, ook de waarschijnlijkheidsdistributie van de andere cijfers in getallen is niet uniform. Dit verschijnsel treedt niet uitsluitend op in het decimale talstelsel, maar treedt op in alle talstelsels.

Ook de kansverdeling van het tweede en de volgende cijfers is niet uniform. De kans dat een getal bijvoorbeeld begint met de cijfers 1, 2 en 3 is ongelijk aan de kans dat een getal begint met de cijfers 1, 2 en 8.

Hoewel het verschijnsel ook in de numerieke wiskunde aandacht trekt – zie Buchanan (1992) – is dit fenomeen in het bijzonder bekend bij oudere technici en ingenieurs, uit het tijdperk van vóór de elektronische rekenmachines, die nog met rekenlinialen hebben leren rekenen, dus de ingenieurs die voor ongeveer 1975 hun opleiding hebben gevolgd.

Het idee van ten opzichte van elkaar verschuivende logaritmische schaalverdelingen (zie figuur 1), dat het feitelijke basisprincipe van de rekenliniaal is, werd in 1624 bedacht door William Oughtred (1575-1660; figuur 2). Merk op, dat de schalen C en D op de liniaal in figuur 1 gewone logaritmische schaalverdelingen met grondtal 10 zijn.

De rekenliniaal is 350 jaar lang het standaardrekeninstrument geweest waarmee snel en doeltreffend eenvoudige tot tamelijk ingewikkelde berekeningen konden worden uitgevoerd.

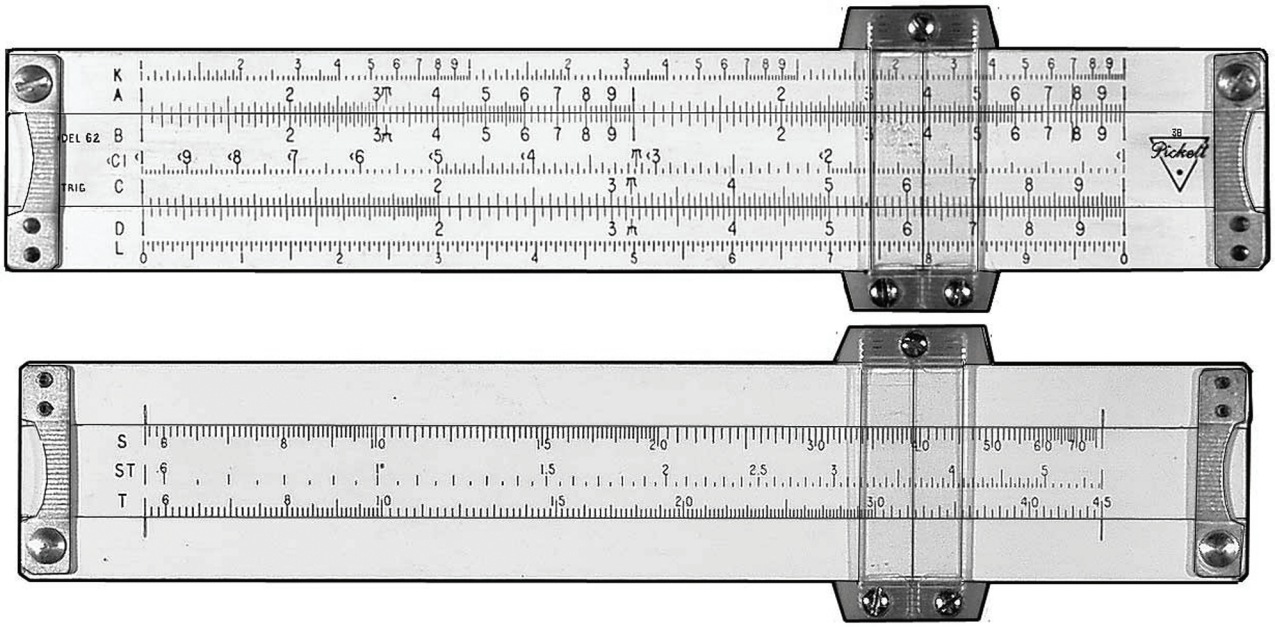


fig. 1 een elementaire rekenliniaal, systeem Rietz

Rond het jaar 1975 kwamen de eerste betaalbare elektronische rekenmachines op de markt en een kleine tien jaar later de eerste kleine en betaalbare computers, maar voor die tijd werden veel berekeningen, die geen al te grote nauwkeurigheid (maximaal drie, soms vier significante cijfers) vereisten, uitgevoerd op rekenlinialen met algemeen inzetbare rekensystemen, zoals Mannheim, Rietz en Darmstadt. Dit gold vooral voor de technische vakgebieden, zoals elektrotechniek en civiele techniek, maar ook voor de economie, landmeetkunde en scheikunde. Daarnaast werden, voor een enorme verscheidenheid aan toepassingen, gespecialiseerde rekenlinialen ontwikkeld.

In het Nederlandse VO-wiskundeonderwijs werden in de jaren zeventig voornamelijk rekenlinialen volgens het systeem Rietz gebruikt; studenten aan bijvoorbeeld HTS'en en technische universiteiten gebruikten linialen, die naast de gebruikelijke schalen volgens één van de hierboven genoemde systemen, ook vakspecifieke schalen bevatten. Zo gebruikten de studenten in de elektrotechniek speciale elektrorekenlinialen.

Een belangrijk kenmerk van het rekenen op rekenlinialen is dat bijna alle berekeningen in één logaritmische decade worden uitgevoerd. Zie bijvoorbeeld schaal C in figuur 1. Alle getallen worden dus voorgesteld in de wetenschappelijke notatie met een mantisse van maximaal drie of vier (significante) cijfers. De rekenaar moet de bijbehorende exponent van 10 apart berekenen; in de praktijk komt dat meestal neer op het mentaal bijhouden van de juiste plaats van de decimale komma in het eindresultaat van de berekeningen.

Hoewel het rekenen op een liniaal het hierna volgende iets duidelijker laat zien, is echter ook zonder rekenliniaal het karakter van de volgende transformatie eenvoudig te begrijpen. We gaan uit van een grote verzameling numerieke data D, bijvoorbeeld de lengtes van alle rivieren op de wereld of de bedragen die op de belastingaangiftes voor de inkomstenbelasting van 2007 in Nederland staan.

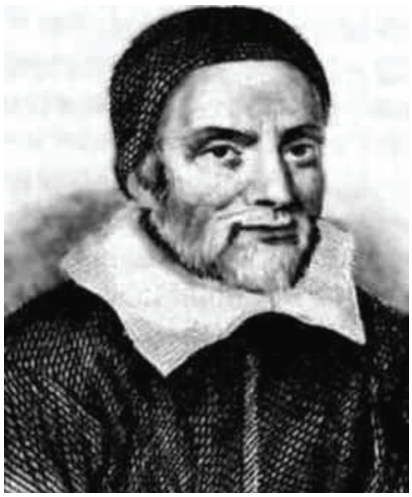


fig. 2 William Oughtred[1]

Ieder getal in de verzameling D vermenigvuldigen we met bijvoorbeeld 2, waarbij een verzameling D' ontstaat van producten. Een getal in D dat met een 1 begint, zal, na vermenigvuldiging met 2, een getal in D' opleveren dat met een 2 of een 3 begint; een getal in D, dat met een 2 begint, zal overgaan in een getal in D' dat met een 4 of een 5 begint; een getal in D, dat 3 als leidend cijfer heeft, zal, na vermenigvuldiging met 2, in D' een getal opleveren dat met een 6 of een 7 begint; een getal in D, dat met het cijfer 4 begint, zal een getal in D' opleveren, dat met een 8 of een 9 begint; een getal in D, dat 5, 6, 7, 8 of 9 als leidend cijfer heeft, zal worden afgebeeld op een getal in D', dat met een 1 begint. Zie tabel 1.

Tabel 1: Verandering van leidend cijfer in vermenigvuldigen met 2

Verandering van leidend cijfer onder transformatie $x \to 2x$	
Leidend cijfer voor transformatie	Leidend cijfer na transformatie
1	2 of 3
2	4 of 5
3	6 of 7
4	8 of 9
5, 6, 7, 8, 9	1

Vaak wordt verondersteld dat het leidende cijfer van de getallen in iedere numerieke dataverzameling uniform is gedistribueerd. Er wordt verondersteld dat de kans dat een getal met het cijfer d begint, gelijk is aan $\frac{1}{9}$, maar een eenvoudige transformatie als de hierbovenstaande laat zien dat die veronderstelling niet juist kan zijn. De kansverdeling van het leidende cijfer in D' moet immers dezelfde zijn als die in D, en dat is onmogelijk bij een uniforme kansverdeling. In veel grote dataverzamelingen komt het cijfer 1 als leidend cijfer veel vaker voor dan het cijfer 9; de kansen voor de andere leidende cijfers liggen ergens tussen de kans voor 1 en de kans voor 9.

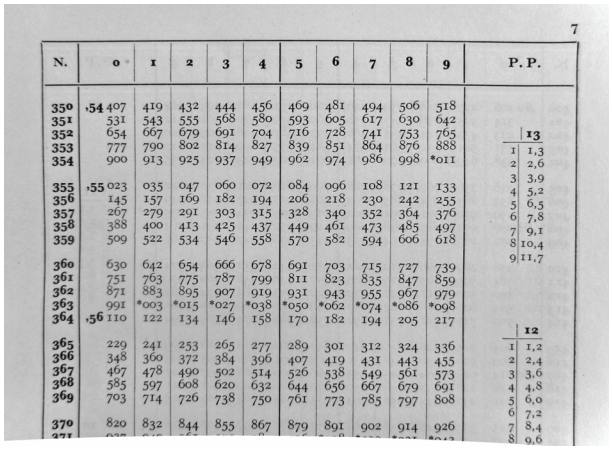
Onder de juiste voorwaarden gehoorzaamt het leidende cijfer van getallen in grote numerieke dataverzamelingen aan een zeer bepaalde universele waarschijnlijkheidsdistributie: als een niet-triviale transformatie $T: D \to D'$ op de dataverzameling wordt uitgevoerd, zoals de bovenstaande vermenigvuldiging met 2, dan is de kansverdeling van het leidende cijfer in de verzameling D' dezelfde als de kansverdeling in D. Dit verschijnsel wordt gewoonlijk de Wet van Benford genoemd, hoewel Benford (1883-1948) (zie figuur 3) niet de eigenlijke ontdekker was. De vraag is aan welke universele kansverdeling het leidende cijfer noodzakelijk voldoet en waarom.

Alle bekende wiskundige verklaringen vooronderstellen een specifieke hypothese. De meest gebruikte hypothese

is de voor de hand liggende schaal-invariantie-hypothese en ook in dit artikel zal de juistheid van die hypothese vooronderstelt worden.

fig. 3 Frank Benford, 1912[2]

Niet alle numerieke dataverzamelingen gehoorzamen echter aan de Wet van Benford. Verzamelingen van toevalsgetallen bijvoorbeeld, die worden gegenereerd door computerprogramma's, of sommige verzamelingen van getallen die aan de normale kansverdeling onderhevig zijn, voldoen niet aan de Wet van Benford. De Wet van Benford lijkt op te gaan voor grote dataverzamelingen, waarin een beperkte mate van toevalligheid optreedt. Voorbeelden zijn in het bijzonder te vinden in grote verzamelingen financiële data. Dat levert belastingdiensten een handig instrument om financiële fraude op het spoor te komen. De Amerikaanse wiskundige Mark Nigrini houdt zich bezig met de implementatie van algoritmen in financiële computerprogramma's, waarin de Wet van Benford wordt toegepast. Zie http://www.aicpa.org/pubs/jofa/may1999/nigrini.htm.



N.	0	1	2	3	4	5	6	7	8	9	P. P.
350	,54 407	419	432	444	456	469	481	494	506	518	
351	531	543	555	568	580	593	605	617	630	642	
352	654	667	679	691	704	716	728	741	753	765	13
353	777	790	802	814	827	839	851	864	876	888	1 1,3
354	900	913	925	937	949	962	974	986	998	*011	2 2,6
355	,55 023	035	047	060	072	084	096	108	121	133	3 3,9
356	145	157	169	182	194	206	218	230	242	255	4 5,2
357	267	279	291	303	315	328	340	352	364	376	5 6,5
358	388	400	413	425	437	449	461	473	485	497	6 7,8
359	509	522	534	546	558	570	582	594	606	618	7 9,1
											8 10,4
360	630	642	654	666	678	691	703	715	727	739	9 11,7
361	751	763	775	787	799	811	823	835	847	859	
362	871	883	895	907	919	931	943	955	967	979	
363	991	*003	*015	*027	*038	*050	*062	*074	*086	*098	
364	,56 110	122	134	146	158	170	182	194	205	217	
											12
365	229	241	253	265	277	289	301	312	324	336	1 1,2
366	348	360	372	384	396	407	419	431	443	455	2 2,4
367	467	478	490	502	514	526	538	549	561	573	3 3,6
368	585	597	608	620	632	644	656	667	679	691	4 4,8
369	703	714	726	738	750	761	773	785	797	808	5 6,0
											6 7,2
370	820	832	844	855	867	879	891	902	914	926	7 8,4
											8 9,6

fig. 4 een halve pagina uit een logaritmetafel, afkomstig uit Noordhoffs wiskundige tafels in 5 decimalen (1953)

Een logaritmische distributie lijkt de universele kansverdeling te zijn voor het leidende cijfer, maar ook voor de andere cijfers in getallen, die een aannemelijke verklaring biedt voor de Wet van Benford.

De formule van Newcomb-Benford

Tot de komst van goedkope elektronische alternatieven in de jaren zeventig werd er niet alleen veel gerekend met rekenlinialen, maar werden ook logaritmetabellen veel toegepast (figuur 4).

In 1881 deed de Amerikaanse wiskundige en astronoom Simon Newcomb (1835-1909) (figuur 5) een curieuze ontdekking: de eerste bladzijden van de tabellenboeken waren overduidelijk veel meer gebruikt dan de laatste bladzijden. Newcomb veronderstelde schertsend dat zijn studenten kennelijk zeer door de tabellenboeken werden aangetrokken, de eerste bladzijden ook ijverig hadden bestudeerd, maar het leek hem dat de studenten al gauw genoeg hadden van het lezen ervan, zoals bij goedkope romannetjes, waarvan het duidelijk is dat verder lezen een verspilling van tijd en moeite is. Newcomb realiseerde zich dat fysici, astronomen en ingenieurs meer te maken hadden met getallen die met de cijfers 1 of 2 beginnen, dan met getallen die met een 8 of een 9 beginnen. Dit verschijnsel kan bij uitstek worden opgemerkt bij het rekenen op rekenlinialen: de linkerhelften van de basisschalen C en D (zie figuur 1), worden veel vaker gebruikt dan de rechterhelften. Newcomb ontdekte een eenvoudige formule, waarmee de kans op het voorkomen van het cijfer *d* als leidend cijfer kan worden berekend:

$$P(leidend\ cijfer = d) = {}^{10}\log\left(1 + \frac{1}{d}\right) \qquad (1)$$

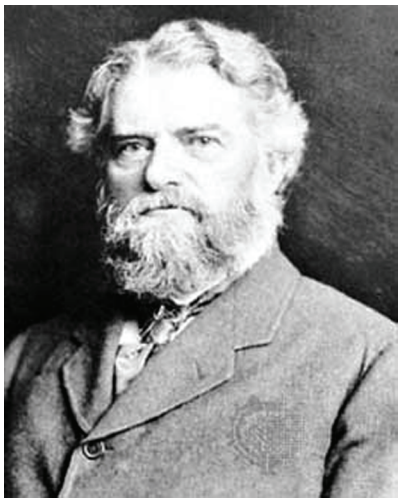


fig. 5 Simon Newcomb[3]

Figuur 6 toont deze discrete kansverdeling.

Omdat hij geen mathematisch of statistisch bewijs leverde voor de juistheid, is het niet bekend hoe Newcomb op

het idee van deze formule kwam. In zijn tijd werden rekenlinialen veel toegepast bij het maken van berekeningen, het ligt dus voor de hand te veronderstellen dat hij via de logaritmische schaalverdeling van de basisschalen (C en D) zijn formule heeft opgesteld. Deze formule geldt overigens niet uitsluitend voor de schaalverdelingen van rekenlinialen, maar voor elke logaritmische schaalverdeling.

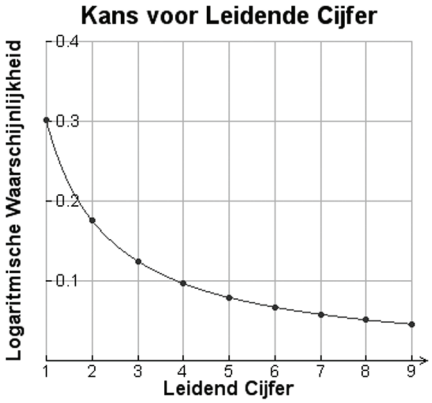


fig. 6 continue grafiek en de discrete kansverdeling volgens formule (1)

Voor logaritmische schaalverdelingen, met grondtal 10, geldt immers voor de afstand tussen twee opeenvolgende gehele getallen d en $d + 1$:

$$D(d, d+1) \propto {}^{10}\log(d+1) - {}^{10}\log(d) = {}^{10}\log\left(1 + \frac{1}{d}\right) \quad (2)$$

Frank Benford herontdekte formule (1) in 1938 (Benford, 1938) en sinds die tijd spreekt men over de Wet van Benford, maar dit fenomeen wordt ook wel *first digit law* of *leading digit phenomenon* genoemd. Gezien de geschiedenis van de ontdekking ervan zullen we formule (1) de formule van Newcomb-Benford noemen.

Het eerste cijfer, en zoals we zullen zien, ook de volgende cijfers in getallen, voldoen kennelijk niet aan een uniforme kansverdeling, zoals vaak wordt verondersteld. De kansverdeling blijkt een logaritmische te zijn. In figuur 7 duiden de donkere staven op deze logaritmische distributie van het leidende cijfer.

Benford onderzocht enorme hoeveelheden numerieke data, zoals de lengte en oppervlakte van rivieren en landerijen, de hoogtes van bergen, numerieke fenomenen uit de natuurkunde en mathematische tabellen en getallen die hij vond in kranten en andere tijdschriften. Hij vond daarmee een overvloed aan empirische ondersteuning voor zijn formule. Zie http://mathworld.wolfram.com/BenfordsLaw.html voor meer voorbeelden.

Tegenwoordig is het tamelijk eenvoudig de Wet van Benford aan het werk te zien in computerprogramma's en spreadsheets. Zie Buchanan & Turner (1992). Laat bijvoorbeeld een MS Excel-rekenblad duizend producten van vier toevalsgetallen berekenen volgens de opdracht:

(9*ASELECT ( ) + 1)*(9*ASELECT ( ) + 1)*(9*ASELECT ( ) +1)*(9*ASELECT ( )+1)

Als het rekenblad de relatieve frequenties, waarmee de leidende cijfers 1, 2, ..., 9 voorkomen in de producten, bepaalt en afbeeldt, dan krijgt men een soortgelijke tabel met empirische gegevens als in figuur 7.

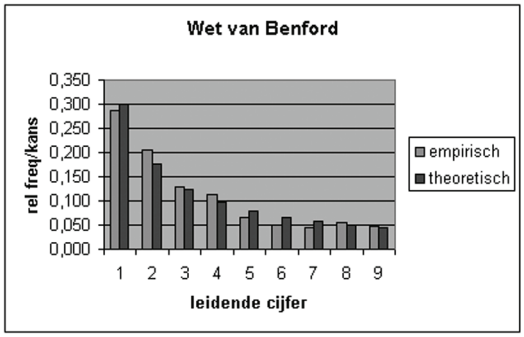


fig. 7 de logaritmische distributie

Tabel 2: Distributie van het leidende cijfer

Leidende Cijfer	Empirische Relatieve Frequentie	Theoretische Kans
1	0,321	0,301
2	0,173	0,176
3	0,116	0,125
4	0,098	0,097
5	0,085	0,079
6	0,062	0,067
7	0,057	0,058
8	0,049	0,051
9	0,039	0,046

Iedere keer als het rekenblad de berekening uitvoert, levert dat een negental waarden in de tweede kolom, die slechts weinig afwijken van de getallen die in eerdere pogingen werden gevonden. Er bestaat een opmerkelijk kleine dispersie tussen de getallen die opeenvolgende trials opleveren. De getallen in de tweede kolom wijken slechts weinig af van de getallen in de derde kolom, die de theoretische kansen vertegenwoordigen.

In figuur 7 representeren de lichter gekleurde staven de kansverdeling voor het leidende cijfer in een typische verzameling van geobserveerde data, terwijl de donkere staven de theoretische, logaritmische distributie vertegenwoordigen.

Schaalinvariantie

In 1961 beschreef Pinkham de schaalinvariantiehypothese als mogelijke verklaring van de Wet van Benford. In 1976 publiceerde Ralph A. Raimi een diepgaand artikel over het First Digit Problem in *American Mathematical Monthly* (Raimi, 1976), waarin hij de toentertijd bekendste verklaringen analyseerde. In 1995 publiceerde Theodore P. Hill (1995) een veel bediscussieerd artikel, eveneens in *American Mathematical Monthly*, waarin hij een uitgebreide analyse geeft van het *significant-digit phenomenon*. In zijn artikel gebruikt Hill fundamentele kansrekening, gebaseerd op maattheorie, bij de constructie van een toepasbaar kansmodel. Daarmee bewijst hij dat de logaritmische distributie voor het leidende cijfer de enig mogelijke universele distributie is. Hill bewijst dat als schaalinvariantie optreedt, de Wet van Benford geldt. Waarschijnlijk is de schaalinvariantie, die Hill in zijn artikel uitgebreid bespreekt, de meest gebruikte verklaring voor de Wet van Benford, maar er bestaan ook andere verklaringen. (Zie Hill (1995), Raimi (1976) en http://www.mathpages.com/home/kmath302/kmath302.htm).

Schaalinvariantie van de kansverdeling voor het leidende cijfer in getallen ligt voor de hand, hoewel die eigenschap problematischer is dan men wellicht zou denken. (Zie Hill (1995)). Schaalinvariantie betekent: als de leidende cijfers in waarden van fysische grootheden aan een universele kansverdeling onderhevig zijn, dan moet deze distributie onafhankelijk zijn van de gekozen eenheden. De distributie moet bijvoorbeeld hetzelfde zijn voor waarden die in *SI-eenheden* worden uitgedrukt als voor waarden die in het systeem van *Imperial Units* worden uitgedrukt.

Als alle getallen in een grote numerieke dataverzameling worden vermenigvuldigd met een constante C, dan impliceert schaalinvariantie dat de kansverdeling van het leidende cijfer in de verzameling veelvouden identiek is aan de kansverdeling voor het leidende cijfer in de oorspronkelijke verzameling. Men krijgt een duidelijker beeld van de logaritmische distributie (1) als men getallen uitdrukt in (decimale) drijvende kommanotatie:

$$x = \pm m \cdot 10^{\pm E} \tag{3}$$

Bijvoorbeeld, zonder vermelding van eenheden, de Newtoniaanse constante van de gravitatie $6,6743 \cdot 10^{-11}$ of de constante van Avogadro $6,0221415 \cdot 10^{23}$.
In (3) is sprake van het grondtal 10, maar mutatis mutandis geldt de Wet van Benford ook in plaatswaardesystemen met een ander grondtal.

Iedere (positieve) mantisse m in (3) ligt in het halfopen interval [1, 10), met als gevolg dat de kansverdeling van het leidende cijfer in de mantissen de universele kansverdeling is voor alle leidende cijfers van alle reële (decimaal voorgestelde) getallen. We zullen zien dat de formule van Newcomb-Benford volgt uit de hypothese van schaalinvariantie.

Alleen functies van de gedaante

$$f(x) = \frac{A}{x};\ x > 0;\ A \neq 0 \tag{4}$$

hebben de eigenschap van schaalinvariantie. Zoals zal blijken, is de waarde van de multiplicatieve constante A van geen belang. A kan dus de waarde 1 krijgen, zodat in het vervolg

$$f(x) = \frac{1}{x} \tag{5}$$

De oppervlakte van een hyperbooltrapezium boven het interval $[a,b]$ en onder de grafiek van $f(x)$ wordt aangeduid met $H(a, b)$. Als ieder getal in het interval $[a,b]$ wordt vermenigvuldigd met een positieve reële constante C, dan is de oppervlakte $H(Ca, Cb)$, van het hyperbooltrapezium boven het interval [Ca,Cb] gelijk aan de eerste oppervlakte:

$$H(Ca, Cb) = H(a, b) \tag{6}$$

In figuur 8 hebben beide hyperbooltrapezia onder de grafiek van $f(x) = \frac{1}{x}$ dezelfde oppervlakte: $H(3, 6) = H(1, 2)$. In dit voorbeeld is de multiplicatieve constante $C = 3$.

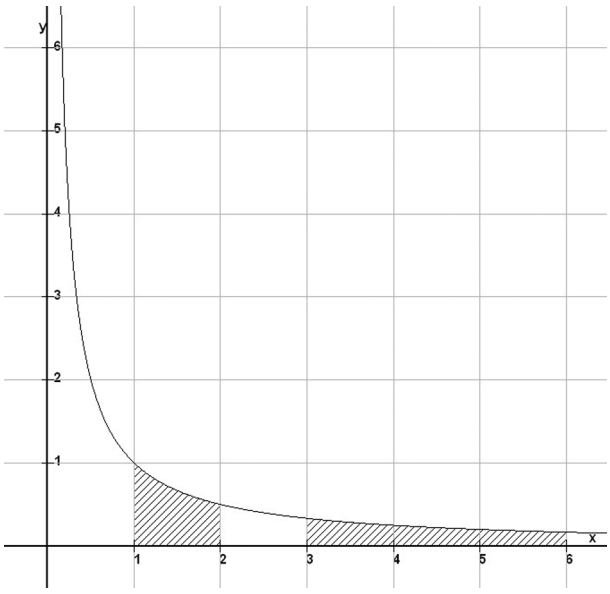


fig. 8 hyperbooltrapezia

Dus voldoen hyperbooltrapezia aan een bruikbare schaalinvariantie. Er geldt:

$$\ln(x) := H(1, x) = \int_1^x \frac{1}{t} dt \tag{7}$$

waaruit kan worden afgeleid

$$H(a, b) = \int_a^b \frac{dx}{x} = \ln(b) - \ln(a) = \ln\left(\frac{b}{a}\right) \qquad (8)$$

Dus

$$H(Ca, Cb) = \ln\left(\frac{Cb}{Ca}\right) = \ln\left(\frac{b}{a}\right) \qquad (9)$$

Tussen haakjes: juist vanwege deze logaritmische eigenschap van hyperbooltrapezia werd de natuurlijke logaritme in het verleden ook wel *hyperbolische logaritme* genoemd.

Als gevolg van de schaalinvariantie van de natuurlijke logaritme van quotiënten (hetzelfde geldt echter ook voor logaritmen met een ander grondtal) kan $f(x) = \frac{1}{x}$ worden gedefinieerd als continue kansdichtheidsfunctie voor een schaalinvariante kansverdeling. Het halfopen interval [1, 10) is de uitkomstenruimte in dit kansmodel. De oppervlakte van het hyperbooltrapezium boven het interval [1, 10] is:

$$\int_1^{10} \frac{dx}{x} = \ln(10) \qquad (10)$$

Als $1 \le a < b < 10$, is de kans op de gebeurtenis 'het getal x ligt in het interval $[a,b]$' gelijk aan:

$$P(x \in [a, b]) = \frac{H(a, b)}{H(1, 10)} = \frac{\int_a^b \frac{dx}{x}}{\ln(10)} = \frac{\ln(b) - \ln(a)}{\ln(10)} = {}^{10}\log\left(\frac{b}{a}\right) \qquad (11)$$

Bijgevolg geldt: in het continue kansmodel, dat drager is van de schaalinvariantie-eigenschap, is kans (noodzakelijkerwijs, zoals bewezen kan worden) gedefinieerd als het quotiënt van de oppervlakten van twee hyperbooltrapezia.

Alle mantissen met leidend cijfer d liggen in het halfopen interval $[d, d + 1)$, $(d = 1, 2, \ldots, 9)$, dus kan men uit (11) deduceren dat de kans op de gebeurtenis 'het leidende cijfer in een decimaal geschreven getal x is gelijk aan d' gelijk is aan:

$$P(\textit{leidend cijfer van } x = d) = {}^{10}\log\left(1 + \frac{1}{d}\right) \qquad (12)$$

Deze formule is de formule van Newcomb-Benford (1)!

De Wet van Benford geldt niet voor alle grote verzamelingen van getallen. Als de getallen onderhevig zijn aan een specifieke distributie, zoals bijvoorbeeld normaal verdeelde gewichten, dan voldoet de verdeling van de leidende cijfers in die verzameling niet aan de Wet van Benford. Als een computer herhaaldelijk een specifieke distributie kiest uit een verzameling stochastische variabelen, als die computer vervolgens een getal trekt uit de waardenverzameling van die stochastische variabele, dan voldoet de verzameling van getrokken getallen wel aan de Wet van Benford. In Pfaltzgraff (2006) vinden we een fraai voorbeeld ter illustratie:

$$\begin{cases} For(N, 1, 1000) \\ randInt(1, 10\,000\,000) \to R \\ randInt(1, R) \to R \\ randInt(1, R) \to R \end{cases}$$

De getallen, die door het laatste statement worden geleverd, voldoen aan de Wet van Benford.

Zoals hierboven vermeld, laat hetzelfde idee, geïmplementeerd in een MS Excel-rekenblad eveneens dit opmerkelijke verschijnsel zien. Henk Pfaltzgraff (2006) geeft een interessante afleiding van de formule van Newcomb-Benford, een afleiding gebaseerd op een idee van Henk Tijms (1999).

The General Significant Digit Law (GSDL)

De Wet van Benford is een speciaal geval van de *General Significant Digit Law* (GSDL). De cijfers in een getal worden van links naar rechts genummerd volgens D_1, D_2, ... Hierin is het leidende cijfer D_1 vanzelfsprekend niet nul. Door middel van de GSDL kan men de kans uitrekenen van het voorkomen van k specifieke cijfers in de eerste k plaatsen in de notatie van een getal. Zijn $d_1, d_2, \ldots, d_k$, die k specifieke cijfers, dan geldt:

$$P(D_1 = d_1 \wedge D_2 = d_2 \wedge \ldots \wedge D_k = d_k) = {}_{10}\log\left(1 + \frac{1}{\sum_{i=1}^{k} d_i \cdot 10^{k-i}}\right) \qquad (13)$$

Bijvoorbeeld, de kans dat een getal 2, 7 en 1 als leidende cijfers heeft is, volgens (13) gelijk aan:

$$P(2, 7, 1) = {}^{10}\log\left(1 + \frac{1}{271}\right)$$

Eigenlijk zien we hier weer de formule van Newcomb-Benford. Zie (2). De verklaring van dit fenomeen is tamelijk eenvoudig: de Wet van Benford is ook geldig voor niet-decimale talstelsels, in het bijzonder het (niet erg praktische) talstelsel met het grondtal 1000.

Tabel 3: Distributie van het leidende cijfer

Kansverdeling van het leidende cijfer voor diverse grondtallen									
	Grondtal								
Cijfer	2	3	4	5	6	7	8	9	10
1	1,0000	0,6309	0.5000	0,4307	0,3869	0,3562	0,3333	0.3155	0,3010
2		0,3691	0,2925	0,2519	0,2263	0,2084	0,1950	0,1845	0,1761
3			0,2075	0,1787	0,1606	0,1478	0,1383	0,1309	0,1249
4				0,1386	0,1245	0,1147	0,1073	0,1016	0,0969
5					0,1018	0,0937	0,0877	0,0830	0,0792
6						0,0792	0,0741	0,0702	0,0669
7							0,0642	0,0608	0,0580
8								0,0536	0,0512
9									0,0458
Som =	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

Grondtalinvariantie

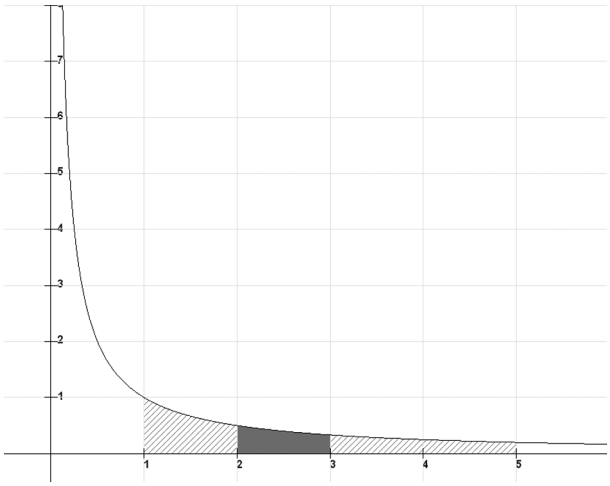


fig. 9 kans als verhouding van twee oppervlakten

Hill (1995) bewijst dat schaalinvariantie grondtalinvariantie impliceert. Dus in een plaatswaardesysteem met grondtal β, waarin getallen worden uitgedrukt, is de kans dat het leidende cijfer D_1 de waarde $d \in \{1, 2, \ldots, \beta - 1\}$ heeft, gelijk aan:

$$P(D_1 = d) = {}^{\beta}\log\left(1 + \frac{1}{d}\right) \qquad (14)$$

Dus bijvoorbeeld, in het plaatswaardesysteem met het grondtal $\beta = 5$, is de kans dat een getal 2 als leidend cijfer heeft:

$$P(x \in [2, 3)) = \frac{H(2, 3)}{H(1, 5)} = \frac{\ln(3) - \ln(2)}{\ln(5)} = {}^{5}\log\left(1 + \frac{1}{2}\right)$$

.

Figuur 9 laat de grafiek zien van $f(x) = \frac{1}{x}$. In figuur 9 is deze kans de verhouding van de donkere oppervlakte $H(2, 3)$ en de totale gemarkeerde oppervlakte $H(1, 5)$.

Tabel 3 toont de kansen op het voorkomen van het leidende cijfer voor verschillende waarden van het grondtal.

Op het eerste gezicht lijkt de keuze voor $f(x) = \frac{A}{x}$ als kansdichtheidsfunctie tamelijk willekeurig te zijn, maar eenvoudig kan worden bewezen dat de logaritmische distributie (13) de unieke continue grondtalonafhankelijke distributie is van de leidende cijfers in grote dataverzamelingen is. Zie Hill, 1995.

In een plaatswaardesysteem met grondtal β gaat formule (13) over in:

$$P(D_1 = d_1 \wedge D_2 = d_2 \wedge \ldots \wedge D_k = d_k) = {}^{\beta}\log\left(1 + \frac{1}{k}\right) \qquad (15)$$

Hier is vanzelfsprekend $d_1 \neq 0$.

Als bijvoorbeeld getallen worden voorgesteld in het binaire talstelsel, dan is de kans dat het binair voorgestelde getal de vier bits 1001 heeft als leidende bits, gelijk aan:

$$P(1,0,0,1) = {}^{2}\log\left(1+\frac{1}{1\cdot 2^3+0\cdot 2^2+0\cdot 2^1+1\cdot 2^0}\right) =$$

$${}^{2}\log\left(1+\frac{1}{9}\right) \approx 0,1520$$

Uit (14) kan met weinig moeite de volgende formule worden afgeleid. De kans dat het *k*-de cijfer in een getal, dat wordt voorgesteld in een plaatswaardestelsel met grondtal β, de specifieke waarde *d* heeft, is gelijk aan:

$$P(D_k = d) = \sum P(D_1 = d_1 \wedge D_2 = d_2 \wedge \ldots \wedge D_{k-1} = d_{k-1} \wedge D_k = d) \quad (16)$$

In (16) loopt de sommatie over alle mogelijke waarden van $d_1, d_2, \ldots, d_{k-1}$. Uit (16) leidt men af:

$$P(D_k = d) = \sum_{j=\beta^{k-2}}^{\beta^{k-1}-1} {}^{\beta}\log\left(1+\frac{1}{j\beta+d}\right) \quad (17)$$

Het volgende voorbeeld illustreert formule (16). Volgens (16) is in het decimale talstelsel de kans dat het derde cijfer een 5 is, gelijk aan:

$$P(D_3 = 5) = \sum_{d_2=1}^{9}\sum_{d_1=0}^{9} P(D_1 = d_1 \wedge D_2 = d_2 \wedge D_3 = 5) \quad (18)$$

Wegens (13) betekent dit:

$$P(D_3 = 5) = \sum_{d_2=1}^{9}\sum_{d_1=0}^{9} {}^{10}\log\left(1+\frac{1}{d_2\cdot 10^2+d_1\cdot 10^1+5}\right) \quad (19)$$

In de noemer lopen de getallen $d_2\cdot 10^1+d_1\cdot 10^0$ van 10 tot en met 99. Omdat het derde cijfer 5 is, moeten we deze getallen nog één plaats naar links schuiven, door middel van een vermenigvuldiging met 10.
Als gevolg daarvan doorloopt de noemer
$(d_2\cdot 10^1+d_1\cdot 10^0)\cdot 10+5 = d_2\cdot 10^2+d_1\cdot 10^1+5$

de getallen 105 tot en met 995. Formule (19) leidt tot:

$$P(D_3 = 5) = \sum_{j=10}^{99} {}^{10}\log\left(1+\frac{1}{j\cdot 10+5}\right) \approx 0,098788 \quad (20)$$

Merk op, dat de telvariabele *j* loopt van $10 = 10^{3-2}$ tot en met $99 = 10^{3-1}-1$. Omdat de quotiënten

$\left(\frac{1}{j\cdot 10+5}\right)$ betrekkelijk klein zijn, kan (20) worden

benaderd door

$$P(D_3 = 5) = \frac{1}{\ln(10)}\sum_{j=10}^{99}\left(\frac{1}{j\cdot 10+5}\right) \approx 0,099982 \quad (21)$$

Conclusie: de logaritmische schaal is een kansmodel
De hierboven afgeleide formules zouden het simpele feit kunnen verduisteren dat de logaritmische schaalverdeling een illustratief voorbeeld is van een kansruimte, met lengte als een maat voor de kans. De lengte van de gehele schaal wordt als 1 gedefinieerd. De lengte van een deelinterval is identiek aan de kans dat men de mantisse van een numerieke waarde uit een grote dataset in het betreffende subinterval vindt. Bijvoorbeeld: volgens de formule van Newcomb-Benford (11) is de kans dat een getal het cijfer 5 als leidend cijfer bezit, gelijk aan de lengte van het interval van 5 tot 6 (zie figuur 10). De kans dat een getal begint met de cijfers 5 en 7 is gelijk aan de lengte van het interval tussen 5.7 en 5.8; de kans dat een getal de cijfers 5, 7 and 4 als de drie leidende cijfers heeft, is gelijk aan de lengte van het interval tussen 5.74 en 5.75, enzovoort.

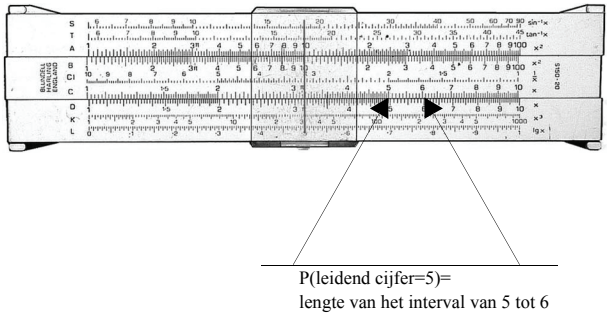


fig. 10 de logaritmische schaal (op een rekenliniaal) als kansmodel

Toepassingen

Wie het lemma 'Benford's Law' intikt in de zoekbalk van een zoekmachine, vindt op internet een enorm aantal verwijzingen. Een belangrijke toepassing van de Wet van Benford blijkt fraudedetectie in grote verzamelingen financiële gegevens te zijn. Het blijkt voor mensen uitermate lastig te zijn om grote verzamelingen getallen te genereren die precies gehoorzamen aan de Wet van Benford. De cijfers op het belastingformulier van een fraudeur bijvoorbeeld blijken anders verdeeld te zijn dan volgens de GSDL zou moeten worden verwacht. Zie de website van AICPA: http://www.aicpa.org/pubs/jofa/may1999/nigrini.htm.

Buchanan en Turner (1992) demonstreren een toepassing van de Wet van Benford in de numerieke analyse. Zij vergelijken de verdeling van representatiefouten in binaire en hexadecimale drijvende kommarepresentaties van reële getallen. Juist door de toepasbaarheid van de Wet van Benford blijkt, in dat verband, het binaire talstelsel superieur te zijn ten opzichte van het hexadecimale.

Dankwoord

De auteur wil hierbij zijn collega Luuk Koens, docent wiskunde aan het Adriaan Roland Holst College, Hilversum, hartelijk bedanken voor het lezen, becommentariëren en verbeteren van eerdere versies van het manuscript. Verder bedankt de auteur Dr. Mark J. Nigrini van Saint Michael's College, Colchester, Vermont, USA, voor zijn toestemming om de foto van Frank Benford te mogen gebruiken. Dr. Nigrini is een specialist op het terrein van toepassingen van de Wet van Benford. Zie zijn website: http://www.nigrini.com/.

Simon van der Salm
Intop Zorgsector, Hilversum

Dit artikel is eerder – met kleine wijzigingen – onder de titel *Benford's Logarithmic Distribution of Digits* verschenen in *The Journal of the Oughtred Society*, Vol. 16, No. 2, 2007, pp. 26-31.

Literatuur

Benford, F. (1938). The law of anomalous numbers, *Proceedings of the American Philosophical Society*, *78*(4), 551-572.

Buchanan, J. L. & Turner, P.R. (1992). *Numerical Methods and Analysis.* McGraw-Hill, Inc, 12-13 en 25-26.

Hill, T.P. (1995). The Significant-Digit Phenomenon. *American Mathematical Monthly*, April 1995, 322-327.

Pfaltzgraff, H. (2006). De Wet van Benford. Tegen de verwachting in beginnen getallen vaker met een 1 dan met een 9. Hoe zit dat? *Euclides,* 81, 301-303.

Raimi, R.A. (1976). The First Digit Problem. *American Mathematical Monthly*, Aug./Sept., 521-537.

Tijms, H. (1999). *Spelen met Kansen*. Utrecht: Epsilon Uitgaven, 44.

Noordhoffs wiskundige tafels in 5 decimalen, vijfde druk (1953).

Noten

[1] http://www-groups.dcs.st-and.ac.uk/Mathematicians/Oughtred.html
[2] foto van: http://www.nigrini.com/Benford's_law.htm
[3] http://turnbull.mcs.st-and.ac.uk/history/Mathematicians/Newcomb.html

---

Wie bedenkt de mooiste opgave?

Op zoek naar gegevens over de kerstboom in de zendmast van IJsselstein, de omslagfoto, belandden we op de site www.degrootstekerstboom.nl. Zo weten we dat sinds 12 december de lampjes weer branden, dat de toren eigenlijk de Gerbrandytoren heet, dat de langste tuidraden 405 m lang zijn, dat de top 375 m hoog is, dat er 120 lampjes van slechts 35 Watt gebruikt worden, enzovoorts. Bij het zien van al deze gegevens ontstond meteen de gedachte, maar dat kan vakdeformatie zijn, dat je daar heel mooie wiskundeopdrachten van kunt maken. En de tweede gedachte: laten we er een soort wedstrijd van maken...

Wie bedenkt de mooiste opgave over de grootste kerstboom ter wereld?
Stuur uw opdrachten naar wiskrant@fi.uu.nl. De beste opdrachten komen in het decembernummer van 2009 en de maker wordt beloond met een eigen lampje in de boom. Vrolijk kerstfeest!