Benford's Law

Vuk Glisovic
3829464

18 augustus 2014

Inhoud

1 Introductie

Voordat we beginnen met waar de wet van Benford vandaan komt en waar deze terugkomt, zullen we eerst bekijken wat deze wet eigenlijk zegt. Deze wet wordt ook wel de significante-cijfer-wet genoemd. De kracht van deze wet is dat het bij vele empirische data gebruikt kan worden. Intuïtief gezien, zou je zeggen dat de proportie waarin het eerste significante cijfer van de getallen uit empirische data voorkomt, $\frac{1}{9}$ is. Wat echter blijkt, is dat veel empirische data juist voldoet aan de wet van Benford! De wet gaat als volgt voor het eerste significante cijfer:

$$P(\text{eerste significante getal} = d) = \log_{10}(1 + \frac{1}{d}),\ d = 1, 2, ..., 9 \tag{1}$$

In algemene vorm, waarin ook het tweede significante getal en meer terugkomen, geldt het volgende: laat hierin $D_i,\ i \geq 1$ (de base 10) significante-cijfer functies zijn zodat bijvoorbeeld $D_1(0,00549) = 1$, $D_2(0,00549) = 4$ en $D_3(0,00549) = 9$. Dan geldt nu voor alle $d_1 \in \{1, 2, ..., 9\}$ en voor alle $d_j \in \{0, 1, ..., 9\},\ j = 2, 3, ..., k$

$$P(D_1 = d_1, ..., D_k = d_k) = \log_{10}(1 + (\sum_{i=1}^{k} d_i * 10^{k-i})^{-1}) \tag{2}$$

Let op! Hierboven zijn de logaritmes dus van base 10. Ook zullen van nu af aan alle logaritmes van base 10 zijn, tenzij anders vermeld. Deze wet kan namelijk nog verder gegeneraliseerd worden tot elke base groter of gelijk aan 2. Dit zal worden behandeld in sectie 4.

Nu we weten welke wet centraal staat, zullen we in het kort behandelen hoe deze wet is ontstaan en in welke gebieden het reeds gebruikt wordt. In 1881 ontstonden de eerste tekenen van leven van de wet van Benford. Simon Newcomb was de eerste wie het opviel dat wanneer hij in een boek met logaritmische tabellen (tabellen met standaardlogaritmen) bladerde, de pagina's waarbij het eerste significante cijfer van de uitkomst een 1 waren, zichtbaar meer versleten waren dan de andere pagina's. Na wat heuristieken, kwam hij tot dezelfde conclusie als hierboven. Newcomb had echter noch een domein of betekenis, noch een formeel argument of numerieke data voor zijn conclusie. Desondanks wordt daarom tegenwoordig ook vaak verwezen naar de wet van Newcomb-Benford in plaats van de wet van Benford.

Circa 57 jaar later heeft de Amerikaan Frank Benford de wet (her)ontdekt en heeft sterk empirisch bewijs geleverd voor deze stelling. Met ruim 20.000 observaties uit 20 verschillende gebieden heeft hij zijn wet krachtig onderbouwd. Hij gebruikte data zo divers als bijvoorbeeld de lengte van 335 rivieren, de warmtecapaciteit van 1389 chemische verbindingen (warmtecapaciteit is de hoeveelheid hitte nodig om de temperatuur van 1 gram chemische verbinding met 1 graad te laten stijgen) en honkbalstatistieken. Sinds Benford zijn bevindingen heeft getoond aan de wereld, zijn er nog veel meer voorbeelden bijgekomen waarin de wet van Benford terugkomt. Een paar leuke en interessante illustraties hiervan zullen nu volgen. Knuth (1969) en Burke en Kincanon (1991) merkten op dat

van de meest gebruikte natuurkundige constantes, zoals lichtsnelheid en zwaartekracht, ongeveer 30% het cijfer 1 als leidend getal heeft. Buck, Merchant en Perez (1993) observeerden dat onder 477 radioactieve halfwaardetijden (van α-straling) de frequentie van het eerste significante cijfer van beide gemeten en berekende waardes, de wet van Benford goed benadert. Nigrini en Wood (1995) lieten zien dat de volks-telling in 1990 van de 3141 plaatsen in de verenigde staten de wet van Benford zeer dicht volgt.

Een mooie toepassing van de wet van Benford, is dat deze wordt gebruikt om fraude op te sporen. De wet van Benford wordt veelvuldig gebruikt (met succes) in de zakenwereld en administratie waar veel boekhouding en belastingheffing voorkomt en waar dus ook veel met getallen gemanipuleerd kan worden. 100 recent gepubliceerde artikelen in twee economische tijdschriften zijn op de proef gesteld met de eerste-significante-cijfer-wet van Benford. 25 van deze artikelen, bleek niet te voldoen an de wet van Benford, wat veel meer is dan je kan verwachten bij onvervalste steekproeven. Dit zegt niet direct dat er fraude is gepleegd en in het algemeen is het ook niet per se zo dat wanneer je niet voldoet aan de wet van Benford, je fraude pleegt. Echter deze methode is wel een snelle en eenvoudige methode om de eerste sporen van fraude te detecteren. Hier kan vervolgens natuurlijk op worden ingespeeld door dat specifieke artikel onder de loep te nemen.

In dit verslag, zal de algehele setting worden behandeld voor de wet van Benford. Dit zal in sectie 2 worden gedaan. Vervolgens bekijken we in secties 3 en 4 respectievelijk schaalinvariantie en base-invariantie. Dit zijn twee bijzondere eigenschappen van de wet van Benford. Tot slot zullen we in sectie 5 een mooi voorbeeld bekijken van de verkiezingen uit 2004 in de Verenigde Staten en het referendum over de terugtrekking van de president Venezuela ook uit 2004. Ook hier zullen we interessante waarnemingen tegenkomen.

2 Domein Benford's law

In dit hoofdstuk zullen we de wiskundige opzet voor de wet van Benford opstellen. We moeten namelijk eerst de correcte ruimte waarnaar we gaan kijken creëren. De functies/stochasten waar het om draait zijn natuurlijk $D_1, D_2, ...$ zoals in hoofdstuk 1 is gedefiniëerd. Hier moeten we dus een juist kansgebied voor vormen. Ook zullen we al een klein opzetje geven naar hoofdstuk 3.

2.1 Intuïtieve verklaring

Veel mensen denken; het getal 1 komt eerder in ons telsysteem (één, twee, drie, vier,...) dan de andere getallen, dus het is logisch dat de 1 vaker voorkomt als eerste significante getal. Een veelgebruikt beginpunt om een bewijs te leveren voor de wet van Benford, is daarom ook om te beginnen met de natuurlijke getallen $\mathbb{N}$. In deze setting geldt $\{D_1 = 1\} = \{1, 10, 11, 12, 13, ..., 19, 100, 101, ...\}$ en om vervolgens te kijken naar de volgende limiet: $\lim_{n\to\infty} \frac{1}{n}\#(\{D_1 = 1\} \cap \{1, 2, 3, ..., n\})$. Hierin geeft # de cardinaliteit aan van desbetreffende verzameling. Echter, deze limiet bestaat niet, maar oscilleert tussen de waardes $\frac{1}{9}$ en $\frac{5}{9}$. Dit kan je als volgt voor je zien; wanneer $n = 9, 99, 999, ...$ dan krijg je $\frac{1}{9} = \frac{11}{99} = \frac{111}{999} = ...$ en voor $n = 19, 199, 1999, ...$ krijg je de reeks $\frac{11}{19} = 0,5789$, $\frac{111}{199} = 0,5578$, $\frac{1111}{1999} = 0,5556$ welke convergeert naar $\frac{5}{9} = 0,5555...$. Met deze kennis zouden we dus elk punt in het interval $[\frac{1}{9}; \frac{5}{9}]$ kunnen kiezen als de proportie/kans van de limiet.

Deze intuïtieve benadering van het probleem gaat dus niet de wiskundige onderbouwing geven die we zoeken. We zullen het dus op een andere manier aan moeten pakken. In de volgende subsecties zullen we de andere aanpak beschrijven en vervolgens in hoofdstukken 3 en 4 respectievelijk schaalinvariantie en base-invariantie bewijzen.

2.2 Significant cijfer vs de significand

De functies $D_1, D_2, ...$ is waar alles om draait. Zodra wij deze netjes in een kansruimte kunnen beschrijven, hebben we al een groot deel van het werk gedaan. Om een breder beeld van deze functies te vormen, bekijken we twee voorbeelden. $D_1 \in \{1, 2, ..., 9\}$ en $D_m \in \{0, 1, ..., 9\}$ voor $m \geq 2$ aangezien het eerste significante cijfer niet nul kan zijn. Ook geldt bijvoorbeeld $\{x \in \mathbb{R} \mid D_1(x) = 1\} = \bigcup_{n=-\infty}^{\infty}[1, 2) * 10^n$ waarin we alle intervallen hebben gestopt in $\mathbb{R}^+$ waarvoor geldt dat voor elk getal uit dat interval, het eerste significante cijfer de 1 is. Tevens geldt $\{x \in \mathbb{R} \mid D_2(x) = 1\} = (\bigcup_{n=-\infty}^{\infty}[11, 12) * 10^n) \bigcup (\bigcup_{n=-\infty}^{\infty}[21, 22) * 10^n) \bigcup ... \bigcup (\bigcup_{n=-\infty}^{\infty}[91, 92) * 10^n)$ waarin dus hetzelfde principe is gebruikt als bij het eerste voorbeeld, we hebben alle intervallen gepakt waarvoor elk willekeurige getal geldt dat het tweede significante cijfer een 1 is. Je ziet dat dit al snel ingewikkelder wordt naarmate m in D_m stijgt. We moeten voor deze functies een juist σ-algebra zien te vormen. Een σ-algebra is in het kort een familie van deelverzamelingen, hier komen we zodirect nog formeel op terug.

Nu zullen we kijken naar de significand. De significand is eigenlijk de coëfficiënt van wanneer een getal in wetenschappelijke notatie wordt gezet. Een eenvoudig voorbeeld is het getal 759, dan is de wetenschappelijke notatie $7,59 * 10^2$ waarin $7,59$ de coëfficiënt is waar we naar zoeken. We definiëren de significand functie $S : \mathbb{R} \to [1, 10)$ als volgt: Voor $x \neq 0$ is $S(x) = t$ met t het unieke getal in $[1, 10)$ met $|x| = t * 10^k$ voor een $k \in \mathbb{Z}$ en $S(0) = 0$ voor het gemak. We kunnen S ook letterlijk geven door

$$S(x) = 10^{\log|x| - \lfloor \log|x| \rfloor} \text{ voor alle } x \neq 0$$

Als we dan wederom 759 als voorbeeld nemen zien we: $S(759) = 10^{\log|759| - \lfloor \log|759| \rfloor} = 10^{\log(7,59)+\log(100)-\lfloor \log(7,59)+\log(100) \rfloor} = 10^{\log(7,59)+2-\lfloor \log(7,59)+2 \rfloor} = 10^{\log(7,59)+2-2} = 7,59$. Deze functie S zal later nog heel belangrijk worden, onthoud dus goed wat deze doet.

We kunnen nu een duidelijke connectie tussen het significante cijfer en de significand aanduiden als volgt:

$$S(x) = \sum_{m=1}^{\infty} 10^{1-m} D_m(x) \tag{3}$$

$$D_m(x) = \lfloor 10^{m-1} S(x) \rfloor - 10 \lfloor 10^{m-2} S(x) \rfloor \text{ voor elke } m \in \mathbb{N} \tag{4}$$

Ofwel, we kunnen dus S volledig in termen van de significante cijfer functies $D_1, D_2, ...$ schrijven en we kunnen elke significante cijferfunctie D_m schrijven in termen van de significand functie S.

2.3 De Kansruimte

Om een correct bewijs te leveren voor de wet van Benford, moeten we dit wel doen met de juiste kanstheorie uitvoeren. Daarom moeten we eerst een aantal formaliteiten opstellen wat we in dit onderdeel zullen doen. Hiermee zullen we een exacte kansruimte definiëren voor de wet van Benford opstellen.

Ten eerste is het belangrijk de juiste kansruimte op te stellen, maar voordat we deze opstellen, moeten we eerst weten wat een kansruimte is en hoe deze werkt. Een kansruimte is $(\Omega, \mathcal{A}, \mathbb{P})$ waarin Ω een niet-lege verzameling is, $\mathcal{A}$ is a σ-algebra op Ω en $\mathbb{P}$ is een kansmaat op $(\Omega, \mathcal{A})$. De definitie van een σ-algebra $\mathcal{A}$ is als volgt:

- $\varnothing \in \mathcal{A}$ en $\Omega \in \mathcal{A}$
- als $A \in \mathcal{A}$ dan $A^c \in \mathcal{A}$
- als $A_n \in \mathcal{A}$ voor alle $n \in \mathbb{N}$ dan $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{A}$

Dus de lege verzameling en Ω zitten er in, alle complementen zitten in $\mathcal{A}$ en alle (af)telbare verenigingen zitten er ook in. σ-algebra's kunnen worden gegenereerd door collecties van verzamelingen en door functies. We gaan nu beide mogelijkheden bekijken. Laat $\mathcal{E}$ een collectie van deelverzamelingen van Ω zijn.

Dan noemen we $\sigma(\mathcal{E})$ het σ-algebra op Ω gegenereerd door $\mathcal{E}$. Dit is de doorsnede van alle σ-algebras die $\mathcal{E}$ bevatten. Ofwel, het kleinste σ-algebra welke $\mathcal{E}$ bevat. Eén van de meest belangrijke σ-algebras is de Borel-σ-algebra $\mathcal{B}$ op $\mathbb{R}$. Eenvoudig gezegd is $\mathcal{B}$ de σ-algebra gegenereerd door alle intervallen. Dus alle verenigingen en complementen van deze intervallen zijn ook bevat in $\mathcal{B}$ volgens de definitie van een σ-algebra. We zeggen ook dat $\mathcal{B}(E) = \{E \bigcap B | B \in \mathcal{B}\}$ met $E \subset \mathbb{R}$. Dus $\mathcal{B}(E)$ is de σ-algebra gerestricteerd tot de verzameling E. De Borelverzameling zal later ook nog veel gebruikt worden in dit verslag.

Tevens kunnen σ-algebras dus door functies worden gegenereerd, dit gaat als volgt; laat $f : \Omega \to \mathbb{R}$ een functie zijn, dan noemen we $\sigma(f)$ het σ-algebra gegenereerd door f. Dat wil zeggen $\sigma(f)$ is de kleinste σ-algebra op Ω welke alle verzamelingen van de vorm $\{\omega \in \Omega \mid a \leq f(\omega) \leq b\}$ voor elke $a \leq b \in \mathbb{R}$ bevat. Een voorbeeld hiervoor is een bernoulli stochast X op $(\mathbb{R}, \mathcal{B})$ die de waardes 0 en 1 aan kan nemen, geeft het volgende σ-algebra:

$$\sigma(X) = \{\varnothing, \{0\}, \{1\}, \{0,1\}, \mathbb{R}\backslash\{0\}, \mathbb{R}\backslash\{1\}, \mathbb{R}\backslash\{0,1\}, \mathbb{R}\}$$

Dit genereren van een σ-algebra kan ook gegeneraliseerd worden naar een familie van functies (wat ook nodig zal zijn i.v.m. $D_1, D_2, ...$ wat een familie van functies is). Stel nu dat $\mathcal{F}$ een familie van functies $f_i : \Omega \to \mathbb{R}$ voor $i \in \{1, 2, ..., n\}$ is, dan geldt

$$\sigma(\mathcal{F}) = \sigma(\bigcup_{i \in \{1,2,...,n\}} \sigma(f_i))$$

wat dus de kleinste σ-algebra is welke alle verzamelingen van de vorm $\{\omega \in \Omega \mid a \leq f_i(\omega) \leq b\}$ voor alle $a \leq b \in \mathbb{R}$ en alle $f_i \in \mathcal{F}$ bevat. Dit zal later nog duidelijker en intuïtiever worden wanneer we dit gaan gebruiken in combinatie met $D_1, D_2, ...$ de significante cijfer functies. Ook is het handig om nog het volledig origineel van E onder een functie $f : \Omega \to \mathbb{R}$ duidelijk te definiëren als

$$f^{-1}(E) = \{\omega \in \Omega \mid f(\omega) \in E\} \text{ voor elke } E \subset \mathbb{R}$$

Tot slot nog een belangrijk onderdeel van de kansruimte, namelijk de kansmaat $\mathbb{P}$. Er geldt dat $\mathcal{A}$ het domein en $[0,1]$ het bereik van $\mathbb{P}$ is, dus $\mathbb{P} : \mathcal{A} \to [0,1]$. Eigenschappen van $\mathbb{P}$ zijn $\mathbb{P}(\varnothing) = 0$, $\mathbb{P}(\Omega) = 1$ en als $A_n \in \mathcal{A}$ disjunct zijn voor alle $n \in \mathbb{N}$, dan geldt $\mathbb{P}(\bigcup_{n\in\mathbb{N}} A_n) = \sum_{n\in\mathbb{N}} \mathbb{P}(A_n)$.

Een voorbeeld van een kansmaat is de uniforme verdeling genoteerd als $\lambda_{a,b}$ op $([a,b), \mathcal{B}([a,b)))$ waarin dus $[a,b) = \Omega$ en $\mathcal{B}([a,b))$ de σ-algebra is. Deze wordt gegeven door

$$\lambda_{a,b}([c,d]) := \frac{d-c}{b-a} \text{ voor elke } [c,d] \subset [a,b)$$

Het speciale geval waarin $b - a = 1$, wordt de Lebesque maat genoemd. Deze Lebesgue maat is een belangrijke maat die later zijn nut zal tonen tijdens het bewijs voor schaalinvariantie. Daar zal in het bijzonder $a = 0$ en $b = 1$ worden gebruikt.

Het is nogmaals belangrijk om de juiste kansruimte te identificeren voor de wet van Benford. Dit houdt dus in dat we in het bijzonder een correcte σ-algebra moeten creëren. We zullen snel zien dat er maar één specifiek σ-algebra is welke ideaal is en waar we dus naar zoeken. Om het al een beetje te verklappen, het is een sub-σ-algebra van $\mathcal{B}$ (de Borel σ-algebra).

We definiëren nu de significand σ-algebra $\mathcal{S}$ op $\mathbb{R}^+$ welke wordt gegenereerd door de significand functie S. Dat wil dus zeggen $\mathcal{S} = \mathbb{R}^+ \bigcap \sigma(S)$. Dus $\mathcal{S}$ bevat alle verzamelingen van de vorm $\{\omega \in \mathbb{R}^+ \mid a \leq S(\omega) \leq b\}$ voor elke $a, b \in \mathbb{R}$. Een voorbeeld van een verzameling welke in $\mathcal{S}$ voorkomt is $\bigcup_{n=-\infty}^{\infty} [1,2) * 10^n = ... \bigcup [0.1,\ 0.2) \bigcup [1,2) \bigcup [10,20) \bigcup [100,200) \bigcup ...$ ofwel alle getallen $x \in \mathbb{R}^+$ waarvoor het eerste significante getal van x een 1 is.

De volgende stelling zal laten zien dat de significand σ-algebra $\mathcal{S}$ de σ-algebra is waar we naar moeten kijken en niet de Borel σ-algebra. Het zegt namelijk dat alle verzamelingen in $\mathcal{S}$ beschreven kunnen worden in termen van de significand functie of equivalent daaraan de significante cijfer functies $D_1, D_2, ...$

Stelling 2.1

(a) Voor elke $A \in \mathcal{S}$ geldt:

$$A = \bigcup_{k \in \mathbb{Z}} 10^k S(A) \text{ met } S(A) = \{S(x) \mid x \in A\} \subset [1,10)$$

(b) $\mathcal{S} = \mathbb{R}^+ \bigcap \sigma(D_1, D_2, ...) = \{\bigcup_{k \in \mathbb{Z}} 10^k B \mid B \in \mathcal{B}([1,10))\}$

bewijs
Deel (a); we weten

$$\mathcal{S} = \mathbb{R}^+ \bigcap \sigma(S) = \mathbb{R}^+ \bigcap \{S^{-1}(B) \mid B \in \mathcal{B}\} = \mathbb{R}^+ \bigcap \{S^{-1}(B) \mid B \in \mathcal{B}([1,10))\}$$

waarin de eerste twee gelijkheden per definitie gelden en de derde gelijkheid geldt aangezien het bereik van S $[1,10)$ is. Deze gelijkheden vertellen ons nu dat voor een gegeven $A \in \mathcal{S}$ er een $B \in \mathcal{B}([1,10))$ bestaat met $A = \mathbb{R}^+ \bigcap S^{-1}(B) = \{x \in \mathbb{R}^+ \mid S(x) \in B\} = \bigcup_{k \in \mathbb{Z}} 10^k B = \bigcup_{k \in \mathbb{Z}} 10^k S(A)$ en we hebben nu onze identiteit in (a).
Voor deel (b) gebruiken we (3) en (4). Uit (3) kunnen we halen dat S volledig is bepaald door de significante cijfer functies. Dit betekent dat $\sigma(S) \subseteq \sigma(D_1, D_2, ...)$ en dus $\mathcal{S} = \mathbb{R}^+ \bigcap \sigma(S) \subseteq \mathbb{R}^+ \bigcap \sigma(D_1, D_2, ...)$. Vervolgens geldt volgens (4) dat D_m volledig is bepaald door S voor elke $m \in \mathbb{N}$. Hieruit volgt nu dat $\sigma(D_1, D_2, ...) \subseteq \sigma(S)$ en dus $\mathbb{R}^+ \bigcap \sigma(D_1, D_2, ...) \subseteq \mathbb{R}^+ \bigcap \sigma(S) = \mathcal{S}$. Dus de eerste gelijkheid in (b) is nu bewezen ($\mathcal{S} = \mathbb{R}^+ \bigcap \sigma(D_1, D_2, ...)$). De tweede gelijkheid gaan we wederom doen door twee inclusies uit te voeren tussen $\mathcal{S}$ en $\{\bigcup_{k \in \mathbb{Z}} 10^k B \mid B \in \mathcal{B}([1,10))\}$. Neem een verzameling $A \in \mathcal{S}$. We weten dat $S(A) \in \mathcal{B}([1,10))$ en dus, gebruik makend van onderdeel (a), geldt $A = \bigcup_{k \in \mathbb{Z}} 10^k B$ voor $B = S(A)$. Nu volgt hieruit $\mathcal{S} \subseteq \{\bigcup_{k \in \mathbb{Z}} 10^k B \mid B \in \mathcal{B}([1,10))\}$. Neem anderzijds een verzameling $\bigcup_{k \in \mathbb{Z}} 10^k B$ voor een willekeurige

$B \in \mathcal{B}([1,10))$. Aangezien $B \in \mathcal{B}([1,10))$ geldt dat $\mathbb{R}^+ \bigcap S^{-1}(B)$ zeker een element is van $\mathcal{S}$. Dit kan je zien door wederom gebruik te maken van onderdeel (a) en $S(A) = B$ te kiezen en dan te zien dat desbetreffende verzameling een element is van $\mathcal{S}$.

□

Wat we zojuist hebben laten zien, is dat het σ-algebra gegenereerd door de significand functie S (het σ-algebra $\mathcal{S}$) of gegenereerd door de familie van functies $D_1, D_2, ...$ $(\sigma(D_1, D_2, ...))$, concreet wordt gegeven door $\{\bigcup_{k\in\mathbb{Z}} 10^k B \mid B \in \mathcal{B}([1,10))\}$. Nu hebben we duidelijk de juiste kansruimte gecreëerd voor de wet van Benford. Sterker nog, we kunnen nu de wet vanuit twee perspectieven benaderen. Namelijk vanuit de significand functie S en vanuit de significante cijfer functies D_m, $m \geq 1$.

Merk op dat $\mathcal{S}$ een sub-σ-algebra van $\mathcal{B}$ is. Dit omdat $\mathcal{B}$ wordt gegenereerd door alle intervallen, terwijl $\mathcal{S}$ bestaat uit verzamelingen wat verenigingen zijn van intervallen. We bekijken hieronder twee voorbeelden om een duidelijk beeld van $\mathcal{S}$ te vormen.

1. $[1,2) \in \mathcal{B}$ aangezien het een interval is. Echter $[1,2) \notin \mathcal{S}$. Dit kan je intuïtief inzien doordat de significant functie D_1 geen onderscheid maakt tussen elementen uit $[1,2)$ en bijvoorbeeld elementen uit $[10,20)$. Daarom komt $[1,2)$ in $\mathcal{S}$ alleen voor in de gedaante van $\bigcup_{k\in\mathbb{Z}} 10^k * [1,2)$

2. De verzameling $\{10^k \mid k \in \mathbb{Z}\} = \{..., 0.01, 0.1, 1, 10, 100, ...\}$ zit in $\mathcal{S}$. Het mooie is, dat we dankzij stelling 2.1 dit op meerdere manieren kunnen inzien. Bijvoorbeeld door op te merken dat deze verzameling equivalent is met $\{x > 0 \mid D_1(x) = 1, D_m(x) = 0 \text{ voor alle } m \geq 2\}$. Of door stelling 2.1(b) en te gebruiken dat $(\{1\}) \in \mathcal{B}([1,10))$ zodat desbetreffende verzameling gelijk is aan $\bigcup_{k\in\mathbb{Z}} 10^k\{1\}$

2.4 Eigenschappen $\mathcal{S}$

In deze subsectie zullen we wat nuttige eigenschappen behandelen van de kansruimte die we in sectie 2.3 hebben gevormd. Daarnaast zullen we het eerste opzetje geven richting de logaritmische verdeling in de wet van Benford.

Om te beginnen gaan we een lemma opstellen waarin we deze eigenschappen gaan behandelen.

Lemma 2.2
De volgende eigenschappen gelden in de significand σ-algebra $\mathcal{S}$:

(a) $\mathcal{S}$ is op zichzelf slaand onder vermenigvuldigingen van gehele machten van 10. Dat wil zeggen

$$10^k A = A \text{ voor elke } A \in \mathcal{S}, k \in \mathbb{Z}$$

(b) $\mathcal{S}$ is gesloten onder vermenigvuldiging met een scalar. Dat wil zeggen

$$\alpha A \in \mathcal{S} \text{ voor elke } A \in \mathcal{S}, \alpha > 0$$

(c) $\mathcal{S}$ is gesloten onder machten van de vorm $\frac{1}{n}, n \in \mathbb{N}$. Dat wil zeggen

$$A^{\frac{1}{n}} \in \mathcal{S} \text{ voor elke } A \in \mathcal{S}, n \in \mathbb{N}$$

bewijs
Deel (a) volgt eenvoudig uit stelling 2.1(a) aangezien $S(10^n A) = S(A)$. Dit betekent niks meer dan het verplaatsen van de komma verandert niks aan de significante cijfers.
Neem voor onderdeel (b) een willekeurige $A \in \mathcal{S}$. Met stelling 2.1(b) weten we dat er een $B \in \mathcal{B}([1,10))$ zodat $A = \bigcup_{k\in\mathbb{Z}} 10^k B$. We nemen een willekeurige $1 < \alpha < 10$. Dit kunnen we doen, omdat stel dat α buiten de gegeven grenzen valt, we α in wetenschappelijke notatie kunnen schrijven met behulp van factoren van 10 zodat de coëfficiënt van α binnen de grenzen 1 en 10 valt. In onderdeel (a), hebben we net gezien, vallen de factoren van 10 weg en houden we dus de coëfficiënt van α over. Veronderstel dus nogmaals zonder verlies van algemeenheid dat $1 < \alpha < 10$. Nu geldt

$$\alpha A = \bigcup_{k\in\mathbb{Z}} 10^k \alpha B = \bigcup_{k\in\mathbb{Z}} 10^k ((\alpha B \bigcap [\alpha, 10)) \bigcup (\frac{\alpha}{10} B \bigcap [1, \alpha))) = \bigcup_{k\in\mathbb{Z}} 10^k C$$

Hierin is $C = (\alpha B \bigcap [\alpha, 10)) \bigcup (\frac{\alpha}{10} B \bigcap [1, \alpha)) \in \mathcal{B}([1,10))$ en dus is $\alpha A \in \mathcal{S}$ wat geslotenheid onder vermenigvuldiging van een scalar aantoont.
Laten we nog even snel kijken naar hoe $C = (\alpha B \bigcap [\alpha, 10)) \bigcup (\frac{\alpha}{10} B \bigcap [1, \alpha))$ is ontstaan. Hier kunnen we het beste een voorbeeld voor gebruiken. Neem $\widetilde{B} = [3, 9)$ en $\alpha = 2$. Dan geldt $2\widetilde{B} = [6, 18)$ wat zich terugvertaalt in $\widetilde{C}$ als volgt: $\widetilde{C} = (2\widetilde{B} \bigcap [3, 10)) \bigcup (\frac{2}{10} \widetilde{B} \bigcap [1, 2)) = ([6, 18) \bigcap [3, 10)) \bigcup ([0.6,\ 1.8) \bigcap [1, 2)) = [6, 10) \bigcup [1,\ 1.8)$ wat dus een element is van $\mathcal{B}$. Als we even het interval $2\widetilde{B} = [6, 18)$ observeren, zien we dat $D_1 \in \{1, 6, 7, 8, 9\}$ en $D_2 \in \{0, 1, ..., 7, 8\}$ als $D_1 = 1$ en $D_m \in \{0, 1, ..., 8, 9\}$ als $D_1 \in \{6, 7, 8, 9\}$ wat ook precies terugvertaald wordt in $\widetilde{C}$.
Tot slot kijken we naar (c). Deze is het lastigst, maar ook zeker te doen. Voordat we beginnen, merken we op dat $\mathcal{B}([1,10))$ wordt gegenereerd door de intervallen van de vorm $[1, t]$ met $1 \leq t < 10$. We weten dat $\mathcal{B}([1,10))$ wordt gegenereerd door alle intervallen in $[1, 10)$. Neem nu een willekeurig interval $[a, b]$ met $a \leq b \in [1, 10)$. Deze kan je vormen door $[1, b] \backslash [1, a]$. Dus de intervallen van de vorm $[1, t]$ met $1 \leq t < 10$ genereren inderdaad $\mathcal{B}([1,10))$ en er geldt $\mathcal{B}([1,10)) = \sigma(\{[1, t] \mid 1 \leq t < 10\})$. Dit betekent dat voor het bewijs van (c) we ons kunnen beperken tot het speciale geval $A = \bigcup_{k\in\mathbb{Z}} 10^k [1, 10^s]$ met $0 < s < 1$. We zoeken wederom naar iets van de vorm $A^{\frac{1}{n}} = \bigcup_{k\in\mathbb{Z}} 10^k C$ met $C \in \mathcal{B}([1,10))$. Voor A geldt:

$$A^{\frac{1}{n}} = \bigcup_{k\in\mathbb{Z}} 10^{\frac{k}{n}} [1, 10^{\frac{s}{n}}] = \bigcup_{k\in\mathbb{Z}} 10^k \bigcup_{j=0}^{n-1} [10^{\frac{j}{n}}, 10^{\frac{j+s}{n}}] = \bigcup_{k\in\mathbb{Z}} 10^k C$$

met $C = \bigcup_{j=0}^{n-1}[10^{\frac{j}{n}}, 10^{\frac{j+s}{n}}] \in \mathcal{B}([1,10))$. Laten we nog even de tweede gelijkheid kort onder de loep nemen. Wat er eigenlijk gebeurt is dat we elke fractie van machten apart behandelen binnen $\bigcup_{j=0}^{n-1}[10^{\frac{j}{n}}, 10^{\frac{j+s}{n}}]$. Merk tevens op dat $10^{\frac{j}{n}} * 10^{\frac{s}{n}} = 10^{\frac{j+s}{n}}$ wat gebruikt is in deze tweede gelijkheid. Uit het bovenstaande volgt dat $A^{\frac{1}{n}} \in \mathcal{S}$ wat geslotenheid onder machten van de vorm $\frac{1}{n}, n \in \mathbb{N}$ aantoont.

□

Merk op dat $\mathcal{S}$ niet gesloten is onder machtnemen van gehele getallen groter dan één. Een eenvoudig voorbeeld hiervoor is $S = \{D_1 = 1\} = \bigcup_{k\in\mathbb{Z}} 10^k[1,2)$. Nu is $S^2 = \bigcup_{k\in\mathbb{Z}} 10^{2k}[1,4) \notin \mathcal{S}$ aangezien bijvoorbeeld interval $[10,40)$ niet bevat is in S^2. Dit is in tegenspraak met bijvoorbeeld lemma 2.2(a). Er geldt namelijk dat $S^2 = \bigcup_{k\in\mathbb{Z}} 10^{2k}[1,4) \neq \bigcup_{k\in\mathbb{Z}} 10^{2k+1}[1,4) = 10 * S^2$. De tegenspraak kan ook worden afgeleid uit stelling 2.1(a).

Tot slot zullen we in dit hoofdstuk nog een zeer nuttig lemma behandelen waarmee we een heel sterke en handige basis hebben om de significand σ-algebra $\mathcal{S}$ te benaderen. We zullen deze namelijk vertalen naar de veelgebruikte Borel σ-algebra op $[0,1)$, dat wil zeggen dat we naar de kansruimte $([0,1), \mathcal{B}([0,1)))$ willen. Voordat we beginnen met het lemma, moeten we eerst nog een kansmaat definiëren. Neem hiervoor kansmaat $\mathbb{P}$ op $(\Omega, \mathcal{A})$ en neem een willekeurige functie $f : \Omega \to \mathbb{R}$ met $\sigma(f) \subseteq \mathcal{A}$ en welke meetbaar is met betrekking tot σ-algebra $\mathcal{A}$. Een functie f is meetbaar met betrekking tot een σ-algebra $\mathcal{G}$ als het volledig origineel van elk interval in $\mathcal{G}$ zit. In formele termen betekent dit het volgende: $\{\omega \in \Omega \mid a \leq f(\omega) \leq b\} \in \mathcal{F}$ voor elke $a \leq b \in \mathbb{R}$. $\mathbb{P}$ en f induceren samen een kansmaat op de volgende manier:

$$f_*\mathbb{P}(B) = \mathbb{P}(f^{-1}(B)) \text{ voor alle } B \in \mathcal{B} \tag{5}$$

Onthoud $f^{-1} : \mathbb{R} \to \Omega$ en $\mathbb{P} : \Omega \to [0,1]$ dus bovenstaande kansmaat is correct gedefiniëerd. Nu specifiek voor ons significand σ-algebra willen we werken met $(\Omega, \mathcal{A}) = (\mathbb{R}^+, \mathcal{S})$ en $f = \log(S)$.

Lemma 2.3
De operatie l_* met functie $l : \mathbb{R}^+ \to [0,1)$ gedefiniëerd door $l(x) = \log(S(x))$ vormt een bijectieve overeenkomst tussen kansmaten op $(\mathbb{R}^+, \mathcal{S})$ en op $([0,1), \mathcal{B}([0,1)))$.

bewijs
Merk eerst op dat de inverse van $l(x)$ gelijk is aan $S^{-1}(10^x)$ voor alle $0 \leq a < b < 1$ aangezien $l(l^{-1}(x)) = l(S^{-1}(10^x)) = \log(S(S^{-1}(10^x))) = \log(10^x) = x$. Dit houdt in dat geldt $l^{-1}([a,b]) = S^{-1}([10^a, 10^b])$ en hieruit kunnen we concluderen dat

$$\sigma(l) = \bigcup_{a,b\in\mathbb{R}} \{\{\omega \in \mathbb{R}^+ \mid a \leq l(\omega) \leq b\}\} = \{l^{-1}(B) \mid B \in \mathcal{B}([0,1))\} =$$

$$\{S^{-1}(10^B) \mid B \in \mathcal{B}([0,1))\} = \{S^{-1}([c,d)) \mid c \leq d \in [1,10)\} = \mathbb{R}^+ \bigcap \sigma(S) = \mathcal{S}$$

waarin $10^B = \{10^s \mid s \in B\}$. Hieruit volgt dat $l_*\mathbb{P}$ een goed gedefiniëerde kansmaat is op $([0,1), \mathcal{B}([1,10)))$ voor elke kansmaat $\mathbb{P}$ gedefiniëerd op $(\mathbb{R}^+, \mathcal{S})$. Dit omdat $l : \mathbb{R}^+ \to [0,1)$ met σ-algebra $\sigma(l) = \mathcal{S}$, $\mathbb{P} : \mathbb{R}^+ \to [0,1)$ met σ-algebra $\mathcal{S}$ en dus $\sigma(l) \subseteq \mathcal{S}(= \mathcal{A})$. Dus het voldoet aan de definitie van (5).

Definiëer X als de verzameling kansmaten op $([0,1), \mathcal{B}([0,1)))$ en Y als de verzameling kansmaten op $(\mathbb{R}^+, \mathcal{S})$. Merk op dat de operatie l_* kansmaten vanuit Y naar kansmaten vanuit X stuurt. Nu, gegeven een kansmaat $\mathcal{P} \in X$ en een $A \in \mathcal{S}$, laat $B \in \mathcal{B}([0,1))$ de unieke verzameling zijn voor welke geldt dat $A = \bigcup_{k\in\mathbb{Z}} 10^k 10^B$ met 10^B gedefiniëerd als hierboven. Merk op dat dit in weze niets anders is dan we gewend zijn. In stelling 2.1 hebben we namelijk gezien dat $A = \bigcup_{k\in\mathbb{Z}} 10^k \widetilde{B}$ met $\widetilde{B} \in \mathcal{B}([1,10))$. We hebben A dus net iets anders geformuleerd. Definiëer tot slot ook nog de functie $g : X \to Y$ door $g(\mathcal{P}) = \mathbb{P}_\mathcal{P}$ met $\mathbb{P}_\mathcal{P}$ de specifieke kansmaat waarvoor geldt $\mathcal{P}(B) = \mathbb{P}_\mathcal{P}(A)$ waarin B en A corresponderen als hierboven. De kansmaat $\mathbb{P}_\mathcal{P}$ is netjes gedefiniëerd op $(\mathbb{R}^+, \mathcal{S})$ aangezien elke $A \in \mathcal{S}$ in $\mathbb{R}^+$ bevat is. Deze identiteit zegt dus eigenlijk dat de kans op $A \in \mathcal{S}$ via kansmaat $\mathbb{P}_\mathcal{P}$ gelijk is aan de kans op $B \in \mathcal{B}([0,1))$ via kansmaat $\mathcal{P}$ voor die unieke B.

We kunnen allereerst eenvoudig inzien dat geldt $l(A) = B$ en $l^{-1}(B) = A$. Er geldt namelijk

$$l(A) = \log(S(A)) = \log(S(\bigcup_{k\in\mathbb{Z}} 10^k 10^B)) = \log(10^B) = B \tag{6}$$

en andersom geldt

$$l^{-1}(B) = S^{-1}(10^B) = \bigcup_{k\in\mathbb{Z}} 10^k 10^B = A \tag{7}$$

We willen nu laten zien dat de operatie $l_* : Y \to X$ een bijectieve operatie is. Dit kunnen we aantonen met behulp van de volgende twee identiteiten, we nemen hiervoor een willekeurige $B \in \mathcal{B}([0,1))$ voor (8) en een (bij B uit (8) horende) $A \in \mathcal{S}$ voor (9):

$$(l_* \circ g)(\mathcal{P}(B)) = l_* g(\mathcal{P}(B)) = l_*\mathbb{P}_\mathcal{P}(B) = \mathbb{P}_\mathcal{P}(l^{-1}B) = \mathbb{P}_\mathcal{P}(A) = \mathcal{P}(B) \tag{8}$$

$$(g \circ l_*)(\mathbb{P}(A)) = g(l_*(\mathbb{P}(A))) = \mathbb{P}_{l_*\mathbb{P}}(A) = l_*\mathbb{P}(B) = \mathbb{P}(l^{-1}(B)) = \mathbb{P}(A) \tag{9}$$

Uit (8) volgt nu dat $(l_* \circ g) = \mathrm{id}_X$ de identiteitsfunctie op X is en uit (9) volgt nu dat $(g \circ l_*) = \mathrm{id}_Y$ de identiteitsfunctie op Y is. Injectiviteit kunnen we nu als volgt aantonen. Laten $y_1, y_2 \in Y$ zijn. Als geldt dat $l_* y_1 = l_* y_2 = x$ met x een kansmaat uit X en $y_1 \neq y_2$, dan zou nu gelden $g(x) = g(l_* y_1) = y_1 \neq y_2 = g(l_* y_2) = g(x)$ wat natuurlijk een tegenspraak is (hierin hebben we (9) gebruikt). Dus $y_1 = y_2$ en hebben we injectiviteit aangetoond.

Vervolgens volgt surjectiviteit uit (8). Deze identiteit houdt namelijk in dat voor elke $x \in X$ een $y \in Y$ is zodat $l_*(y) = x$. Merk op dat $(l_* \circ g) : X \to Y \to X$ wat goed gedefiniëerd is. Tot slot kunnen we nu zeggen dat de correspondentie via l_*; $\mathbb{P} \mapsto l_*\mathbb{P}$ bijectief is.

□

We weten nu dus dat voor elke kansmaat $\mathbb{P}$ op $(\mathbb{R}^+, \mathcal{S})$, we deze over kunnen zetten naar een kansmaat op $([0,1), \mathcal{B}([0,1)))$ met behulp van $l(x) = \log(S(x))$. Deze kansmaat is dan $f_*\mathbb{P}$. Het nut van deze bijectie, zal later bij het bewijs van schaal-/ en base-invariantie aan bod komen.

Merk op dat l als $\log(S)$ is gedefiniëerd en dit feit alleen terugkomt in (6) en (7). We hadden net zo goed de functie $\widetilde{l} = \frac{1}{9}(S(x) - 1)$ kunnen gebruiken in plaats van l. Echter, de l zoals we die in lemma 2.3 hebben gedefiniëerd, heeft een handige relatie met de wet van Benford. Dit zullen we zo hieronder merken, maar om dit in te zien, definiëren we eerst $\mathbb{B}$ als de kansmaat op $(\mathbb{R}^+, \mathcal{S})$ gegeven door

$$\mathbb{B}(\{x > 0 \mid S(x) \leq t\}) = \mathbb{B}(\bigcup_{k\in\mathbb{Z}} 10^k[1,t]) = \log(t) \text{ voor alle } 1 \leq t < 10$$

Dit is eigenlijk de meest natuurlijke formulering van de wet van Benford. Om de connectie met (1) te zien, pakken we het volgende voorbeeld waarin we de kans bekijken dat het eerste significante getal $1, 2$ of 3 is:

$$\log(4) = \mathbb{B}(\bigcup_{k\in\mathbb{Z}} 10^k[1,4)) = \mathbb{B}(\bigcup_{k\in\mathbb{Z}} 10^k[1,2)) + \mathbb{B}(\bigcup_{k\in\mathbb{Z}} 10^k[2,3)) + \mathbb{B}(\bigcup_{k\in\mathbb{Z}} 10^k[3,4)) =$$

$P(\text{eerste significante getal} = 1) + P(\text{eerste significante getal} = 2) + P(\text{eerste significante getal} = 3)$

$= \log(1 + \frac{1}{1}) + \log(1 + \frac{1}{2}) + \log(1 + \frac{1}{3}) = \log(\frac{2}{1} * \frac{3}{2} * \frac{4}{3}) = \log(4)$

In het volgende hoofdstuk zullen we zien dat de Lebesgue maat (uniforme verdeling) op $([0,1), \mathcal{B}([0,1)))$, $\lambda_{0,1}$ genaamd, welke we in sectie 2.3 hebben behandeld, veel nuttige eigenschappen heeft. Er geldt namelijk $l_*\mathbb{B} = \lambda_{0,1}$ met l gedefiniëerd als in lemma 2.3, wat gelijk de relevantie van deze specifieke l aantoont. Dus $l_*\mathbb{B}$ is uniform verdeeld op $([0,1), \mathcal{B}([0,1)))$. Om dit te laten zien, nemen we een willekeurig interval $[a,b]$ met $a \leq b \in [0,1)$, dan nu

$$l_*\mathbb{B}([a,b]) = \mathbb{B}(l^{-1}([a,b])) = \mathbb{B}(S^{-1}(10^{[a,b]})) = \mathbb{B}(\bigcup_{k\in\mathbb{Z}} 10^k 10^{[a,b]}) =$$

$$\mathbb{B}(\bigcup_{k\in\mathbb{Z}} 10^k 10^{[0,b]}) - \mathbb{B}(\bigcup_{k\in\mathbb{Z}} 10^k 10^{[0,a]}) = \log(10^b) - \log(10^a) = \frac{b-a}{1-0}$$

wat precies de uniforme verdeling $\lambda_{0,1}$.

3 Schaalinvariantie

Een zeer bewonderd en interessant fenomeen gerelateerd aan de wet van Benford, is het concept van schaalinvariantie. Eenvoudig gezegd betekent schaalinvariantie dat de wet van Benford moet gelden ongeacht de eenheid waarin wordt gekeken. Dus bijvoorbeeld als een (voldoende grote) dataset de logaritmische verdeling van Benford volgt in meters, dan zou deze nog steeds moeten kloppen in voet (één meter is gelijk aan $3,2808$ maal één voet). Of bijvoorbeeld als we euro's overzetten naar Amerikaanse dollars binnen een dataset, dan zou als in de ene eenheid de wet van Benford geldt, het in de andere eenheid ook moeten gelden. Formeel gesproken houdt dit in; ondanks dat de getallen individueel veranderen, de uitspraken over de algehele verdeling van de significante cijfers zouden niet moeten veranderen.

Sterker nog, de logaritmische verdeling voor significante cijfers in de wet van Benford is de enige verdeling die schaalinvariant is. Dat wil dus ook zeggen dat als een dataset schaalinvariantie vertoont, de dataset voldoet aan de wet van Benford. Dit zal bewezen worden in stelling 3.3. Echter voordat we aan het bewijzen van stelling 3.3 beginnen, moeten we eerst de juiste opzet maken. Dit zullen we doen met behulp van stelling 3.1 en lemma 3.2.

3.1 Opzet tot het bewijs

Zoals we net al zeiden, hebben we voor het bewijs eerst nog een aantal onderdelen nodig. In deze subsectie zullen we wat we nog nodig hebben voor het bewijs van schaalinvariantie behandelen.

Allereerst definiëren we een nieuwe functie $\langle t \rangle$ als het fractionele deel van een getal $t \in \mathbb{R}$. Ofwel formeel gezien: $\langle t \rangle = t - \lfloor t \rfloor$. Hiermee kunnen we de uniforme verdeling modulo 1 voor stochasten en kansmaten definiëren:

- Een stochast X op een kansruimte $(\Omega, \mathcal{A}, \mathbb{P})$ is uniform verdeeld modulo 1 als geldt:
$$P(\langle X \rangle \leq s) = s \text{ voor alle } s \in [0, 1)$$
- Een kansmaat $\mathcal{P}$ op $(\mathbb{R}, \mathcal{B})$ is uniform verdeeld modulo 1 als geldt:
$$\mathcal{P}(\{x \mid \langle x \rangle \leq s\}) = \mathcal{P}(\bigcup_{k \in \mathbb{Z}} [k, k+s]) = s \text{ voor alle } s \in [0, 1)$$

Nu zullen we een stelling gaan bewijzen welke voortborduurt op bovenstaande definities. In eerste instantie lijkt de stelling niet nuttig te zijn voor het concept van schaalinvariantie voor de wet van Benford omdat het over de uniforme verdeling gaat, maar het zal zijn nut zeker waarmaken later in stelling 3.3.

Stelling 3.1
Laten X en Y stochasten zijn, dan geldt:

(a) Als X uniform verdeeld modulo 1 is en Y is onafhankelijk van X, dan is $X + Y$ ook uniform verdeeld modulo 1.

(b) Als $\langle X \rangle$ en $\langle X + \alpha \rangle$ dezelfde verdeling hebben voor een irrationele α, dan is X uniform verdeeld modulo 1.

bewijs
Het bewijs zal worden uitgevoerd met behulp van Fourieranalyse. Voordat we (a) en (b) aan gaan tonen, zullen we eerst de benodigde aspecten van de Fourieranalyse bekijken. Definiëer hiervoor eerst een stochast Z die waardes aanneemt op $\Omega = [0, 1)$ of equivalent hieraan de bijbehorende kansmaat $\mathcal{P}_Z$ op $([0, 1), \mathcal{B}([0, 1)))$. Definiëer tevens:

$$\widehat{\mathcal{P}_Z}(k) = \mathbb{E}(e^{2\pi ikZ}) = \int_{s=0}^{1} e^{2\pi iks}\mathcal{P}_Z(s)ds =$$

$$\int_{s=0}^{1} \cos(2\pi ks)\mathcal{P}_Z(s)ds + i\int_{s=0}^{1} \sin(2\pi ks)\mathcal{P}_Z(s)ds \text{ met } k \in \mathbb{Z}$$

Nu hebben we de oneindige reeks $(\widehat{\mathcal{P}_Z}(k))_{k\in\mathbb{Z}}$ ook wel bekend als de coëfficiënten van Z of van $\mathcal{P}_Z$. Dit is een begrensde reeks met $|\widehat{\mathcal{P}_Z}(k)| \leq 1$ voor alle $k \in \mathbb{Z}$ en $\widehat{\mathcal{P}_Z}(0) = 1$.

Twee zeer belangrijke eigenschappen van Fourier coëfficiënten, zijn dat $(\widehat{\mathcal{P}_Z}(k))_{k\in\mathbb{Z}}$ op unieke wijze $\mathcal{P}_Z$ bepaalt. Dat wil zeggen dat $\mathcal{P}_{Z_1} = \mathcal{P}_{Z_2}$ dan en slechts dan als $(\widehat{\mathcal{P}_{Z_1}}(k)) = (\widehat{\mathcal{P}_{Z_2}}(k))$ voor alle $k \in \mathbb{Z}$. Ten tweede geldt $\widehat{\mathcal{P}_{\langle Z_1+Z_2\rangle}}(k) = \widehat{\mathcal{P}_{Z_1}}(k) * \widehat{\mathcal{P}_{Z_2}}(k)$ voor alle k gegeven dat Z_1 en Z_2 onafhankelijk zijn. Voor een uitgebreide verklaring voor deze twee eigenschappen, is [CT] het juiste boek.

Merk op dat als Z de uniforme verdeling op $[0, 1)$ is (genoteerd als $Z = U([0, 1))$), dat $\widehat{\mathcal{P}_Z}(k)$ eenvoudige coëfficiënten heeft;

$$\widehat{\mathcal{P}_{U([0,1))}}(k) = \begin{cases} 1 & \text{als } k = 0 \\ 0 & \text{anders} \end{cases}$$

Om dit in te zien, merken we eerst op dat de dichtheidsfunctie $\mathcal{P}_U$ van $U([0, 1))$ gegeven wordt door

$$\mathcal{P}_U(s) = \begin{cases} \frac{1}{1-0} = 1 & \text{als } 0 \leq s < 1 \\ 0 & \text{anders} \end{cases}$$

Hiermee volgt het volgende:

$$\widehat{\mathcal{P}_{U([0,1))}}(k) = \int_{s=0}^{1} e^{2\pi iks}\mathcal{P}_Z(s)ds = \int_{s=0}^{1} e^{2\pi iks}ds = \int_{s=0}^{1} 1ds = 1 \text{ als } k = 0$$

en

$$\widehat{\mathcal{P}_{U([0,1))}}(k) = \int_{s=0}^{1} e^{2\pi iks}ds = \frac{1}{2\pi ik}e^{2\pi iks}|_{s=0}^{1} = \frac{1}{2\pi ik}(e^{2\pi ik*1} - e^{2\pi ik*0}) = \frac{1}{2\pi ik}(e^{2\pi ik} - 1)$$

$$= \tfrac{1}{2\pi ik}((\cos(2\pi k) + i\sin(2\pi k)) - 1) = \tfrac{1}{2\pi ik}(1 - 1) = 0 \text{ als } k \in \mathbb{Z}\backslash\{0\}$$

Nu kunnen we beginnen met het bewijzen van (a) en (b). Deze zijn met dit voorbereidende werk niet lang meer.

Deel (a); we weten dat het volgende geldt:

$$\widehat{\mathcal{P}_{\langle X+Y\rangle}}(k) = \widehat{\mathcal{P}_X}(k) * \widehat{\mathcal{P}_Y}(k) = \begin{cases} 1*1=1 & \text{als } k=0 \\ 0 & \text{anders} \end{cases}$$

Merk op dat ongeacht de verdeling van Y, $\widehat{\mathcal{P}_Y}(0) = 1$ (gaat dit zelf na). Bovenstaande vergelijking laat nu zien dat $\langle X+Y\rangle = U([0,1))$. Ofwel dat $X+Y$ uniform verdeeld modulo 1 is (dit kan je vanuit de definitie van uniform verdeeld modulo 1 inzien).

Observeer voor deel (b) dat als we stellen dat $Z = \alpha$ (α is irrationeel) met kans 1, ofwel Z is constant, dan geldt $\widehat{\mathcal{P}_\alpha}(k) = \mathbb{E}(e^{2\pi ik\alpha}) = e^{2\pi ik\alpha}$ voor elke $k \in \mathbb{Z}$. Dit kunnen we gebruiken in het volgende; stel nu dus nog extra dat $\langle X\rangle$ en $\langle X+\alpha\rangle$ dezelfde verdeling hebben. Dan hebben we nu met behulp van de eerste eigenschap die we hebben behandeld voor Fouriercoëfficiënten:

$$\widehat{\mathcal{P}_{\langle X\rangle}}(k) = \widehat{\mathcal{P}_{\langle X+\alpha\rangle}}(k) = \widehat{\mathcal{P}_X}(k) * \widehat{\mathcal{P}_\alpha}(k) = \widehat{\mathcal{P}_X}(k) * e^{2\pi ik\alpha} \text{ voor elke } k \in \mathbb{Z}$$

Als $k = 0$, dan worden we niet veel wijzer van bovenstaande vergelijking, aangezien dan $1 = 1*1$ wat natuurlijk klopt. Bekijk daarom het geval waarin $k \neq 0$. Nu geldt dat $e^{2\pi ik\alpha} \neq 1$ voor alle $k \neq 0$. Stel, integendeel, dat $e^{2\pi ik\alpha} = 1$ voor een $k \neq 0$, bekijk nu: $e^{2\pi ik\alpha} = \cos(2\pi k\alpha) + i\sin(2\pi k\alpha)$. Om dit gelijk aan 1 te krijgen, moet $k\alpha \in \mathbb{Z}$. Nu moet dus $k\alpha = n$ voor een $n \in \mathbb{Z}\backslash\{0\}$ (we weten dat $k\alpha \neq 0$). Echter, dan zou $\alpha = \frac{n}{k} \in \mathbb{Q}$ wat in tegenspraak is met dat α irrationeel is. Dus $e^{2\pi ik\alpha} \neq 1$ voor alle $k \neq 0$. Nu volgt uit bovenstaande vergelijking dat $\widehat{\mathcal{P}_{\langle X\rangle}}(k) = 0$ voor alle $k \neq 0$. Dit houdt nu in dat $\widehat{\mathcal{P}_{\langle X\rangle}}(k) = \widehat{\mathcal{P}_{U([0,1))}}(k)$ en hier volgt uit dat $\langle X\rangle = U([0,1))$. Dus X is uniform verdeeld modulo 1.

□

3.2 Bewijs schaalinvariantie

Voordat we beginnen met het bewijs van schaalinvariantie, moeten we eerst formeel definiëren wat schaalinvariantie inhoudt. Neem hiervoor een σ-algebra $\mathcal{A} \supseteq \mathcal{S}$ op $\mathbb{R}^+$. Een kansmaat $\mathcal{P}$ op $(\mathbb{R}^+, \mathcal{A})$ heeft schaalinvariante significante cijfers als geldt

$$\mathcal{P}(\alpha A) = \mathcal{P}(A) \text{ voor alle } \alpha > 0, A \in \mathcal{S}$$

Met deze definitie kunnen we lemma 3.2 inleiden, welke we nodig hebben voor stelling 3.3.

Lemma 3.2
Laat $\mathcal{P}$ een kansmaat zijn op $(\mathbb{R}^+, \mathcal{A})$ met $\mathcal{A} \supseteq \mathcal{S}$ en zet $\mathcal{Q} := l_*\mathcal{P}$ met l gegeven als in lemma 2.3. De identiteit $\mathcal{P}(\alpha A) = \mathcal{P}(A)$ is nu equivalent aan $\mathcal{Q}(\langle t+B\rangle) = \mathcal{Q}(B)$. Hierin is $\langle t+B\rangle = \{\langle t+x\rangle \mid x \in B\}$, $t = \log(\alpha)$ (dus $t \in \mathbb{R}$ en $\alpha > 0$) en $A = \bigcup_{k\in\mathbb{Z}} 10^k 10^B$.

bewijs
Merk op dat volgens lemma 2.3 $\mathcal{Q}$ nu een kansmaat op $([0,1), \mathcal{B}([0,1)))$ is. Stel dat geldt $\mathcal{P}(\alpha A) = \mathcal{P}(A)$, dan hebben we nu:

$$\mathcal{Q}(B) = \mathcal{P}(l^{-1}(B)) = \mathcal{P}(S^{-1}(10^B)) = \mathcal{P}(\bigcup_{k\in\mathbb{Z}} 10^k 10^B) = \mathcal{P}(A) = \mathcal{P}(\alpha A) =$$

$$\mathcal{P}(\alpha \bigcup_{k\in\mathbb{Z}} 10^k 10^B) = \mathcal{P}(\bigcup_{k\in\mathbb{Z}} 10^k 10^{\log(\alpha)+B}) = \mathcal{P}(\bigcup_{k\in\mathbb{Z}} 10^k 10^{\langle \log(\alpha)+B\rangle}) =$$

$$\mathcal{P}(S^{-1}(10^{\langle t+B\rangle})) = \mathcal{P}(l^{-1}(\langle t+B\rangle)) = \mathcal{Q}(\langle t+B\rangle)$$

Stel anderzijds dat geldt $\mathcal{Q}(\langle t+B\rangle) = \mathcal{Q}(B)$, dan krijgen we nu:

$$\mathcal{P}(A) = \mathcal{P}(\bigcup_{k\in\mathbb{Z}} 10^k 10^B) = \mathcal{P}(S^{-1}(10^B)) = \mathcal{Q}(B) = \mathcal{Q}(\langle t+B\rangle) = \mathcal{P}(l^{-1}(\langle t+B\rangle)) =$$

$$\mathcal{P}(S^{-1}(10^{\langle t+B\rangle})) = \mathcal{P}(\bigcup_{k\in\mathbb{Z}} 10^k 10^{\langle \log(\alpha)+B\rangle}) = \mathcal{P}(\bigcup_{k\in\mathbb{Z}} 10^k 10^{\log(\alpha)+B}) = \mathcal{P}(\alpha A)$$

□

Met deze voorbereiding, kunnen we nu beginnen met stelling 3.3. Dit is dus de stelling die schaalinvariantie aantoont. Merk nogmaals op dat de wet van Benford de enige wet is waarin schaalinvariantie voor kan komen (dit gaan we nu aantonen). Dus als een dataset schaalinvariantie vertoont, dat voldoet het ook aan de wet van Benford.

Stelling 3.3 (Schaalinvariantie)
Een kansmaat $\mathcal{P}$ op $(\mathbb{R}^+, \mathcal{A})$ met $\mathcal{A} \supseteq \mathcal{S}$ heeft schaalinvariante significante cijfers dan en slechts dan als $\mathcal{P}(A) = \mathbb{B}(A)$ voor elke $\mathcal{A} \in \mathcal{S}$ (dat wil zeggen d.e.s.d.a. het voldoet aan de wet van Benford).
bewijs
Merk op dat we de kansmaat $\mathbb{B}$ aan het einde van hoofdstuk 2 hebben behandeld.

Neem een willekeurige, maar vaste, $\mathcal{P}$ op $(\mathbb{R}^+, \mathcal{A})$ en laat $\mathcal{P}_0$ zijn restrictie zijn tot $(\mathbb{R}^+, \mathcal{S})$. Stel tevens $\mathcal{Q} := l_*\mathcal{P}_0$. Merk op dat dit eigenlijk hetzelfde is als in lemma 3.2. Vervolgens weten we ook dankzij lemma 3.2 dat

$$\mathcal{P}_0(\alpha A) = \mathcal{P}_0(A) \text{ voor alle } \alpha > 0, A \in \mathcal{S} \tag{10}$$

nu equivalent is aan

$$\mathcal{Q}(\langle t+B\rangle) = \mathcal{Q}(B)(= \mathcal{Q}(\langle B\rangle)) \text{ voor alle } t \in \mathbb{R}, B \in \mathcal{B}([0,1)) \tag{11}$$

Neem nu een stochast X zó dat zijn verdeling wordt gegeven door $\mathcal{Q}$. Nu betekent (11) niks anders dan dat de verdelingen van $\langle X\rangle$ en $\langle t+X\rangle$ hetzelfde zijn. Met behulp van stelling 3.1(a) en (b) geldt nu dat $\langle X\rangle$ en $\langle t+X\rangle$ dezelfde verdeling hebben dan en slechts dan als X uniform verdeeld modulo 1 is. Stel

namelijk enerzijds dat $\langle X \rangle$ en $\langle t + X \rangle$ dezelfde verdeling hebben, dan volgt uit stelling 3.1(b) dat X uniform verdeeld modulo 1 is. Stel anderzijds dat X uniform verdeeld modulo 1 is, dan is met $Y = t$ (een constante) $t + X$ ook uniform verdeeld modulo 1. Merk op dat een constante altijd onafhankelijk is van elke stochast.

Nu we weten dat X uniform verdeeld modulo 1 moet zijn om te voldoen aan (11), moet dus ook gelden dat $\mathcal{Q} = \lambda_{0,1}$. We hebben nu dus $l_* \mathcal{P}_0 = \mathcal{Q} = \lambda_{0,1} = l_* \mathbb{B}$. De laatste gelijkheid kunnen we halen uit het einde van hoofdstuk 2. Hieruit volgt nu dat $\mathcal{P}_0 = \mathbb{B}$ dan en slechts dan als voldaan wordt aan (10).

□

4 Base-invariantie

Een ander bijzonder fenomeen gerelateerd aan de wet van Benford, is base-invariantie. Dit is de idee dat de wet van Benford ook naadloos over zou moeten gaan naar basen anders dan 10. Het mooie is dat de gebruikte definities, lemma's en stellingen allemaal analoog afgewerkt kunnen worden in elke andere base om dit te bewijzen. Dit resulteert in de volgende verdere generalisatie van (2) van de wet van Benford:

$$P(D_1^{(b)} = d_1, D_2^{(b)} = d_2, ..., D_k^{(b)} = d_k) = \log_b(1 + (\sum_{i=1}^{k} b^{k-i} d_i)^{-1})$$

waarin b de base aangeeft waarin gewerkt wordt.

De hoofdstelling die we in dit hoofdstuk gaan bewijzen, laat zien dat als je in base b zit, je vervolgens over kunt gaan in andere basen welke machten (van de vorm $\frac{1}{n}$ met $n \in \mathbb{N}$) zijn van b. Bekijk, om hier inzicht in te krijgen, het volgende voorbeeld:

Laten we allereerst afspreken dat $S^{(b)}$ de significand functie in base b is. Dus $|x| = b^k S^{(b)}(x)$ waarin $S^{(b)}(x)$ het unieke getal in $[1, b)$ is waarvoor de identiteit geldt (voor een $k \in \mathbb{Z}$).

Laat nu

$$A = \{x > 0 \mid D_1^{(10)} = 1\} = \{x > 0 \mid S^{(10)}(x) \in [1, 2)\},$$

dan is (kan je zelf na gaan) de verzameling

$$A^{\frac{1}{2}} = \{x > 0 \mid S^{(10)}(x) \in [1, \sqrt{2}) \bigcup [\sqrt{10}, \sqrt{20})\}.$$

Nu zullen we ook werken met $S^{(100)}$. Deze is hierboven gedefiniëerd. We kunnen nu inzien dat:

$$A = \{x > 0 \mid S^{(100)}(x) \in [1, 2) \bigcup [10, 20)\}$$

en hieruit volgt nu dat:

$$\{x > 0 \mid S^{(b)}(x) \in [1, b^{\frac{\log_{10}(2)}{2}}) \bigcup [b^{\frac{1}{2}}, b^{\frac{1+\log_{10}(2)}{2}})\} = \begin{cases} A^{\frac{1}{2}} & \text{als } b = 10 \\ A & \text{als } b = 100 \end{cases}$$

Dus als een kansverdeling base-invariantie voor de significante cijfers vertoont op de significand σ-algebra $\mathcal{S}$, dan moet dus gelden dat $P(A) = P(A^{\frac{1}{2}})$. Dit moet ook gelden voor andere machten van de vorm $\frac{1}{n}$ met $n \in \mathbb{N}$. Ofwel, $P(A) = P(A^{\frac{1}{n}})$ moet gelden voor alle $n \in \mathbb{N}$. Merk op dat volgens lemma 2.2(c) $A^{\frac{1}{n}} \in \mathcal{S}$ als $A \in \mathcal{S}$. Dus de kansen waar we naar gaan kijken, zijn goed gedefiniëerd. Nu kunnen we ook base-invariantie definiëren; laat P een kansmaat zijn op $(\mathbb{R}^+, \mathcal{A})$ met $\mathcal{A} \supset \mathcal{S}$, dan heeft deze kansmaat P base-invariante significante cijfers, als $P(A) = P(A^{\frac{1}{n}})$ voor alle $A \in \mathcal{S}$ en $n \in \mathbb{N}$. Deze definitie zegt dus in weze dat als je in een bepaalde base b aan werken bent, je deze base over kan zetten naar een andere base van de vorm $b^{\frac{1}{n}}$ met $n \in \mathbb{N}$. Neem bijvoorbeeld base 10;

We weten dat $1000^{\frac{1}{3}} = 10$ en dus kunnen we vanuit base 10 naar base 1000 overschakelen. Of een tweede voorbeeld; $49^{\frac{1}{2}} = 7$, wat betekent dat we kunnen schakelen tussen deze twee basen.

Merk op dat basen 7 en 10 niks met elkaar te maken hebben aangezien er geen macht van de vorm $\frac{1}{n}$ is waarvoor $10^{\frac{1}{n}} = 7$. Sterker nog de significand σ-algebra's gegenereerd door $S^{(10)}$ en door $S^{(7)}$ zijn geen van beiden bevat in de ander. De hoofdstelling gaat echt puur voor machten van de vorm $\frac{1}{n}$ van een base b met $n \in \mathbb{N}$.

4.1 Bewijs Base-invariantie

Voordat we de hoofdstelling gaan bewijzen, moeten we eerst een aantal functies definiëren. Om te beginnen vormen we δ_a op $(\Omega, \mathcal{A})$ (met een Ω en een σ-algebra) genaamd de Dirac maat door:

$$\delta_a(A) = \begin{cases} 1 & \text{als } a \in A \\ 0 & \text{als } a \notin A \end{cases}$$

Dit is een bijzondere kansmaat aangezien de volledige kans zich focust op één punt. Vanzelfsprekend is dit punt $a \in \Omega$ Vervolgens hebben we de volgende functie $T_n : [0,1) \to [0,1)$ nog nodig. Deze is als volgt gedefiniëerd voor $n \in \mathbb{N}$:

$$T_n(x) = \langle nx \rangle$$

We zeggen ook dat een kansmaat P op $([0,1), \mathcal{B}([0,1))$ T_n-invariant is, als $T_{n*}P = P$. Deze functie en deze laatste definitie zullen binnenkort gebruikt worden.

Voor het bewijs van de hoofdstelling van base-invariantie, hebben we het volgende lemma nodig;

Lemma 4.1
Een kansmaat P op $([0,1), \mathcal{B}([0,1)))$ is T_n-invariant voor alle $n \in \mathbb{N}$ dan en slechts dan als $P = q\delta_0 + (1-q)\lambda_{0,1}$ voor een $q \in [0,1]$.

Een bewijs van dit lemma, kan terug gevonden worden in [BA]. Dit lemma zegt dus in weze dat T_n-invariante kansmaten (voor alle $n \in \mathbb{N}$), geschreven kunnen worden als een convexe combinatie van de Dirac maat geconcentreerd op het punt 0 en de Lebesgue maat $\lambda_{0,1}$. Merk op dat we net als in het bewijs voor schaalinvariantie, desbetreffende kansmaten terug willen vertalen naar de uniforme verdeling. Dit zullen we nu zien.

Stelling 4.2(Base-invariantie)
Een kansmaat $\mathcal{P}$ op $(\mathbb{R}^+, \mathcal{A})$ met $\mathcal{A} \supset \mathcal{S}$ heeft base-invariante significante cijfers dan en slechts dan als voor een $q \in [0,1]$ geldt dat $\mathcal{P}(A) = q\delta_1(A) + (1-q)\mathbb{B}(A)$ voor alle $A \in \mathcal{S}$.
bewijs
Neem een willekeurige kansmaat $\mathcal{P}$ op $(\mathbb{R}^+, \mathcal{A})$ met $\mathcal{A} \supseteq \mathcal{S}$. Noteer door $\mathcal{P}_0$ zijn

restrictie tot $(\mathbb{R}^+, \mathcal{S})$ en laat $\mathcal{Q} = l_*\mathcal{P}_0$ met $l = \log(S)$. Merk op dat dit exact hetzelfde is als in stelling 3.3 (schaalinvariantie).

Een belangrijke stap is om in te zien dat base-invariante significante cijfers voor $\mathcal{P}_0$, equivalent is aan dat $\mathcal{Q}$ T_n-invariant. Dit zullen we nu inzien; neem hiervoor een verzameling $A = \{x > 0 \mid S^{(10)}(x) < 10^s\}$ met uiteraard $0 \leq s < 1$.

$$T_{n*}\mathcal{Q}([0,s)) = \mathcal{Q}(\bigcup_{j=0}^{n-1}[\frac{j}{n}, \frac{j+s}{n})) = l_*\mathcal{P}_0(\bigcup_{j=0}^{n-1}[\frac{j}{n}, \frac{j+s}{n})) = \mathcal{P}_0(S^{-1^{(10)}}\bigcup_{j=0}^{n-1}[10^{\frac{j}{n}}, 10^{\frac{j+s}{n}})) =$$

$$\mathcal{P}_0(\bigcup_{k\in\mathbb{Z}} 10^k \bigcup_{j=0}^{n-1}[10^{\frac{j}{n}}, 10^{\frac{j+s}{n}})) = \mathcal{P}_0(A^{\frac{1}{n}}) = \mathcal{P}_0(A) = \mathcal{P}_0(S^{-1^{(10)}}(10^{[0,s)})) = l_*\mathcal{P}_0([0,s)) = \mathcal{Q}([0,s))$$

Deze redenatie kan ook andersom worden gevolgd, dus met de aanname dat $\mathcal{Q}$ T_n-invariant is.

Nu weten we dus dat $\mathcal{Q}$ T_n-invariant is d.e.s.d.a. $\mathcal{P}_0$ base-invariante significante cijfers heeft. Tevens weten we met behulp van lemma 3.1 dat $\mathcal{Q}$ T_n-invariant is d.e.s.d.a. $\mathcal{Q} = q\delta_0 + (1-q)\lambda_{0,1}$ voor een $q \in [0,1]$.

Merk nu de volgende onderdelen op; laat $B \in \mathcal{B}([0,1))$ zó zijn, dat $A = \bigcup_{k\in\mathbb{Z}} 10^k 10^B$. Dan geldt dat

$$0 \in B \text{ d.e.s.d.a. } 1 \in A.$$

Ten tweede geldt

$$\lambda_{0,1}(B) = \lambda_{0,1}(l(A)) = l_*^{-1}\lambda_{0,1}(A) = \mathbb{B}(A).$$

Uit deze twee onderdelen en $\mathcal{Q} = q\delta_0 + (1-q)\lambda_{0,1}$ volgt dat $\mathcal{P}_0 = q\delta_1(A) + (1-q)\mathbb{B}(A)$ d.e.s.d.a. $\mathcal{P}$ base-invariante significante cijfers heeft.

□

Merk op dat het bewijs hierboven in base 10 is uitgevoerd. Dit kan voor elke willekeurige base groter of gelijk aan 2 analoog uitgevoerd worden.

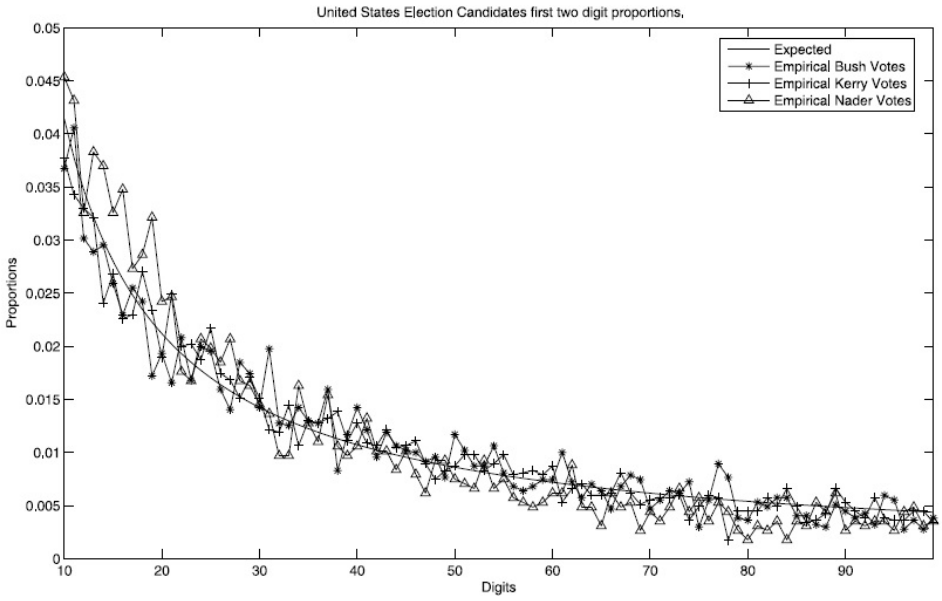
Gevolg 4.3(Base-invariantie)
Een continue kansmaat $\mathcal{P}$ op $\mathbb{R}^+$ heeft base-invariante significante cijfers dan en slechts dan als $\mathcal{P}(A) = \mathbb{B}(A)$ voor alle $A \in \mathcal{S}$ (dat wil zeggen d.e.s.d.a. het voldoet aan de wet van Benford).

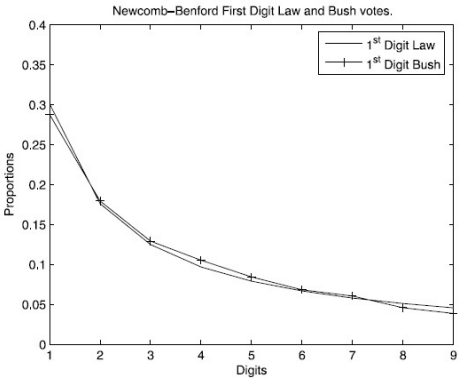
Merk tot slot op dat dus nog een tweede gevolg van base-invariante significante cijfers is, dat als een kansmaat base-invariante significante cijfers heeft, dan heeft deze kansmaat ook schaalinvariante significante cijfers en andersom. Dus als een kansmaat voldoet aan de wet van Benford, heeft deze kansmaat ook schaal-/en base-invariante significante cijfers.

5 Praktijkvoorbeelden

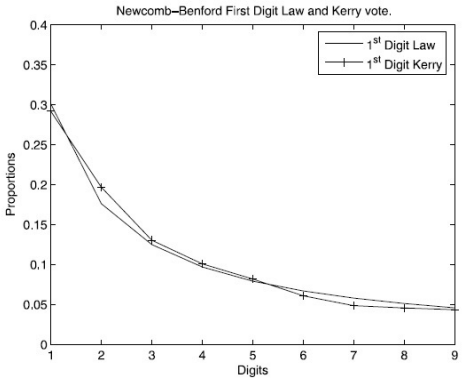
In dit hoofdstuk willen we graag twee mooie praktijkvoorbeelden geven van de wet van Benford. Hiervoor hebben we het artikel [PT] gebruikt. In het eerste voorbeeld gaat het om de Amerikaanse presidentsverkiezingen van 2004, waarin het een close call was tussen Bush en Kerry. We gaan hier kijken naar de eerste twee significante cijfers in meerdere grafieken. Het tweede voorbeeld gaat over het referendum ook in 2004 vóór het aftreden van president Hugo Chávez.

Zoals we zeiden gaan we eerst kijken naar de Amerikaanse verkiezingen van 2004. De data die hier voor is gebruikt, is verkregen uit zo'n 4700 stemlokalen. In elk stemlokaal zijn het aantal stemmen voor Bush, Kerry en Nader geteld en per stemlokaal zijn dus de eerste twee significante cijfers gebruikt van het aantal stemmen per kandidaat. De data voor Bush en Kerry kan je zelf terugvinden op http://us.cnn.com/ELECTION/2004/pages/results/. Dit leverde de volgende figuren op:





Eerste significante cijfer voor Bush



Eerste significante cijfer voor Kerry

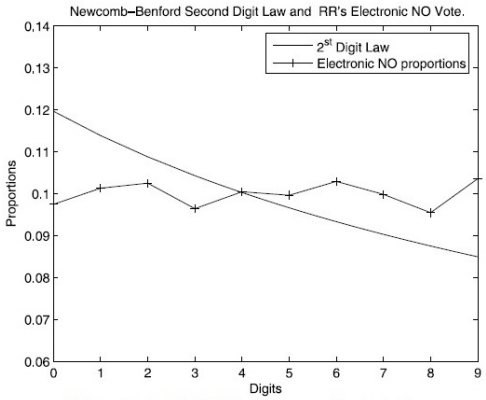
Wat we zien in de eerste figuur, is dat op basis van de eerste twee significante cijfers gecombineerd, we een mooie grafiek krijgen die aardig de logaritmische verdeling van de wet van Benford volgt voor elke van de drie kandidaten.

In de figuren daaronder hebben we gekeken naar het eerste significante cijfer van Bush en Kerry apart. Deze zijn ook heel mooi verdeeld in de figuren. Merk op dat wanneer we kijken naar het tweede significante getal, het cijfer 1 niet op een voorspelling van 30% staat, maar op een voorspelling van ongeveer 12%. Dit kan je zelf nagaan door de volgende formule te gebruiken welke is voortgekomen uit (2) en te sommeren over alle mogelijkheden voor het eerste significante getal:

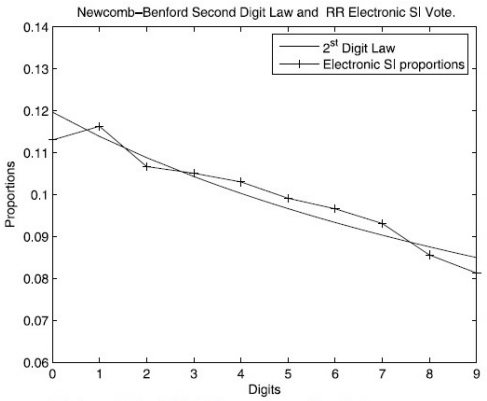
$$\mathcal{P}(D_2 = 1) = \mathcal{P}(D_1 = 1, D_2 = 1) + ... + \mathcal{P}(D_1 = 9, D_2 = 1) = \sum_{k=1}^{9} \log(1 + (10k+1)^{-1})$$

Dit gebruikende, zien we dat in de eerste grafiek de proporties op de verticale as kleiner zijn dan $0,05$ wat ook logisch is wanneer je naar de eerste twee significante cijfers kijkt (en niet naar slechts de eerste).

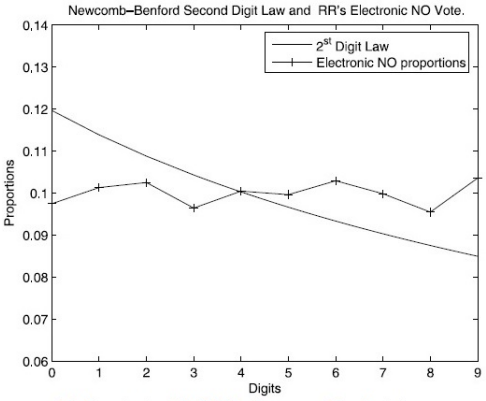
Het tweede voorbeeld waar we naar wilden kijken, was het referendum vóór het aftreden van president Hugo Chávez in 2004. Tijdens deze stemming zijn elektronische stemlokalen en traditionele stemlokalen waar je handmatig met pen en papier je stem uitbracht. De data die is gebruikt, is verkregen uit zo'n 19.000 elektronische stemlokalen en uit ongeveer 4500 traditionele stemlokalen. Hier geldt dat stemmen 'TEGEN' in het voordeel van Hugo Chávez zijn en stemmen 'VOOR' zijn dus in het nadeel van de president. Er is voornamelijk gekeken naar het tweede significante cijfer. Dit leverde de volgende grafieken op:



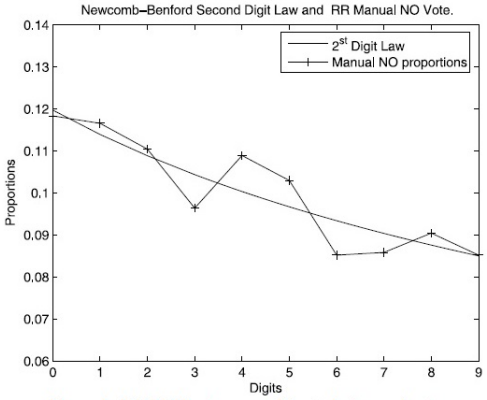
Elektronische 'TEGEN' stemmen. Dus in het voordeel van Hugo CHávez.



Elektronische 'VOOR' stemmen. Dus in het nadeel van Hugo Chávez.



Elektronische 'TEGEN' stemmen. Dus in het voordeel van Hugo CHávez.



Manuele 'TEGEN' stemmen. Dus in het voordeel van Hugo Chávez.

Wat opvalt is dat de elektronische 'TEGEN' stemmen de wet van Benford voor het tweede significante getal niet volgen, maar juist uniform verdeeld zijn. Dit in tegenstelling tot de handmatig uitgebrachte 'TEGEN' stemmen en de elektronische 'VOOR' stemmen! Dit betekent natuurlijk niet direct dat er fraude is gepleegd. Echter, dit is natuurlijk wel heel vreemd dat de wet klopt voor de handmatig uitgebrachte 'TEGEN' stemmen en de elektronische 'VOOR' stemmen, maar niet voor de elektronische 'TEGEN' stemmen. Elektronische stemmen zijn uiteraard eenvoudig te vervalsen. Desalniettemin betekent dit dat alle stemmen onder de loep genomen zouden moeten worden.

6 Nawoord

Om te beginnen vond ik het een leuk en mooi onderwerp om mijn scriptieverslag over te schrijven. Ik vind het mooi om te zien hoe deze theorie toegepast kan worden in de praktijk. Niet alleen vind ik het een fascinerend fenomeen, maar ook ben ik verbaasd over de eenvoud van de uitdrukking voor de significante cijfers. Zeker als je ziet wat er allemaal aan vooraf ging om schaalinvariantie aan te tonen. Ook vond ik het enerverend om te zien dat base-invariantie ook een eigenschap is van de wet van Benford. Het blijft toch bijzonder om te zien dat, ook al zijn we gewend te tellen in base 10 (leuk om te weten dat dit komt omdat de mens 10 vingers heeft), alles overgevoerd kan worden naar elke willekeurige base.

Tevens wil ik mijn dank uiten naar mijn begeleider Tobias Müller wie mij goed geholpen heeft. Niet alleen met bepaalde bewijzen, maar ook met de kleine formaliteiten (de puntjes op de i). Ik hoop dat iedereen heeft genoten van dit verslag en kennis op heeft kunnen doen van de wet van Benford en zijn toepassingen.

7 Bronnen

[BA] MR2846899 (2012h:37015) Reviewed Berger, Arno(3-AB-MS); Hill, Theodore P.(1-GAIT) A basic theory of Benford's law. (English summary) Probab. Surv. 8 (2011), 1126. 37A45 (11K06 60-02 60F15 60G57 62E10)
[Hi1] Hill, T.P. (1995), Base-Invariance Implies Benfords Law, Proc. Amer. Math. Soc. 123(3), 887895. MR1233974
[Hi2] Hill, T.P. (1995), A Statistical Derivation of the Significant-Digit Law, Statis. Sci. 10(4), 354363. MR1421567
[PT] Pericchi, Luis; Torres, David Quick anomaly detection by the Newcomb-Benford law, with applications to electoral processes data from the USA, Puerto Rico and Venezuela. (English summary) Statist. Sci. 26 (2011), no. 4, 502516. MR2951385
[CT] Chow, Y.S. and Teicher, H. (1997), Probability Theory. Indepen- dence, Interchangeability, Martingales (3rd ed.), Springer. MR1476912
[TK] Karl-Heinz Tdter Benford's Law as an Indicator of Fraud in Economics German Economic Review, 2009, vol. 10, pages 339-351.