SJÄLVSTÄNDIGA ARBETEN I MATEMATIK

MATEMATISKA INSTITUTIONEN, STOCKHOLMS UNIVERSITET

En introduktion till Benford's Lag

av

Erik Nyberg

2019 - No K32

MATEMATISKA INSTITUTIONEN, STOCKHOLMS UNIVERSITET, 106 91 STOCKHOLM

En introduktion till Benford's Lag

Erik Nyberg

Självständigt arbete i matematik 15 högskolepoäng, grundnivå

Handledare: Gregory Arone

2019

Sammanfattning:

Innan alla hade en miniräknare till hands, användes böcker innehållandes logaritmtabeller för att beräkna multiplikation av stora tal. En astrolog uppmärksammade något märkligt, att första sidorna i tabellerna var betydligt mer nötta än de sidor som var närmare slutet. Astronomen drog slutsattsen att vi mer frekvent använder oss av tal med lägre första siffra. Denna simpla iaktagelse tog statistiker över hundra år att bevisa, än idag finns det obesvarade frågor om varför så många datamängder följer Benford's lag.

Benford's lag är en universell matematisk lag som hävdar att naturlig data i större utsträckning har en ledade siffra med lågt ett siffervärde. Detta arbete handlar om observerandet av lagen samt förklaringar till varför naturlig data har en tendens följer lagen.

Abstract:

Before everyone had a calculator in their pocket, books containing logarithm tables to calculate multiplication of large numbers. When an astronom suddenly noticed something strange, that the first few pages of the tables were considerably more used than the ones that were nearer the end. The Astronomer concluded that we use frequencies with lower first digits more frequently than a larger one. This simple observation took mathematicians and statisticians over a hundred years to prove, even today there are unanswered questions about Benford's law.

Benford's law is a universal mathematical law that claims that natural data are more likely to have a low significant digit. This report is about observing the law and explaining why natural data tends to obey Benford's law.

Innehållsförteckning

1 Let's play a game!

Låt oss spela ett enkel lek. Det finns två spelare A och spelare B. Spelet går till på följande vis, spelare A väljer ett valfritt tal x och det gör även spelare B som väljer talet y. Sedan multiplicerar vi ihop talen $d = x \times y$. Om den första nollskiljda siffran i $d \in \{1, 2, 3\}$ vinner spelare A, om $d \in \{4, 5, 6, 7, 8, 9\}$ vinner spelare B.

Vem kommer att vinna?

2 Introduktion

Benford lag är ingen intuitiv lag, snarare motsatsen. Man kan jämföra upptäckten av Bedfords lag med upptäcken av Newtons gravitationslag, i hänseende till att lagen uppmärksammades genom simpla observationer. Eftersom Benfords lag endast var en mystisk observation trodde en del matematiker att lagen skulle vara matematiskt obevislig. Värt att notera att det tog 114år från Newcomb publicerade första observationen av Benfords lag tills Theodore P.Hill bevisade lagen för specifika mängder(se 6.2).[Wiki1]

Varför Benford's lag förbryllar många beror på att lagen påstår att om man samlar in större mängd tal tagna ur diverse olika distinkt (tal ur tidningar, fysikaliska konstanter ut physics handbook, avstånd mellan orter osv..), och sedan studerar varje unikt tals ledande siffra (första nollskilda siffran $d_m, m \in \{1, 2, 3, \ldots, 9\}$). Så kommer ni uppmärksamma något förbryllande. Resultatet av våran studie kommer med största sannolikhet resultera i att tal med ett lägre siffervärde på den ledande siffrna kommer vara betydligt vanligare än övriga.

Den vanligaste tanken är att ledande siffran d kommer vara ungefär slumpmässigt fördelat dvs 1/9=11,11 % eftersom det finns nio heltal mellan 1 och 9. Men dessvärre är det så att man med hög sannolikhet kommer uppmärksamma att över 30% av våra insamlade tal kommer att ha en ledande siffra som är en etta jämfört med att chansen för en nia som ledande är mindre än 5%. Så med andra ord hävdar Benford's lag att tal med en etta som ledande siffra, är mer än sex gånger vanligare tal med en nia.

3 Vad är Benfords lag

Benford's lag som bestämmer hur tal naturligt vill fördela sig efter sin ledande siffra. Där Benfrdfördelningen ges av

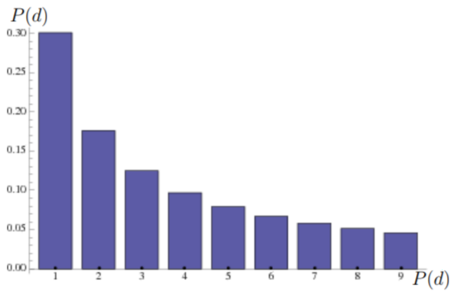
$$P_{BF}(d_m) = \log_{10}(1 + \frac{1}{d_m}), \qquad m \in \{1, 2, 3, .., 9\}$$

Benford's lag,Benford fördelning osv kommer förkortas som BF.

Kärt barn har många namn och mängden författare som skrivit artiklar om BF använder sig matematiska omskrivningar av lagen. De vanligaste är dock följande

$$P(d_m) = \log_{10}(1 + \frac{1}{d_m}) = \log_{10}(\frac{d_{m+1}}{d}) = \log_{10}(d_{m+1}) - \log_{10}(d_m) = \frac{\log(\frac{1+d_m}{d_m})}{\log_{10}}.$$

d	$P(d)$
1	0.3010
2	0.1761
3	0.1249
4	0.0969
5	0.0792
6	0.0669
7	0.0580
8	0.0512
9	0.0458
Sum	1.0000



Illusterar Benford's lag om hur ledande siffran fördelar sig i data som följer BF.[Wol]

Tal som följer BF är finns överallt runtomkring oss det kan vara talen i dagstidningen eller din favorit mattematiska serie. Den första bevisningen av lagens existens gjordes av Benford som sammlade in massor av data tagna från naturen. Bilden här nedan tagen från Benford *The law of anomalous numbers* (1938).

4 Historik

Allt började att astronomen Simon Newcomb 1881 uppmärksammade de första sidorna i dåtidens logaritmtabeller användes mer frekvent än de övriga sidorna. Newcomb insåg att beräkningar med lägre första siffra var vanligare än de övriga, i samband med observationen publicerade Newcomb artikeln *Note on the frequency of use of the different digits in natural numbers"*. Där Newcomb hävdade att den ledande siffran följer fördelningen $P(d) = \log_{10}(d + 1) - \log_{10}(d)$.[Ne]

Till vänster i tabellen kan man se vilket distrik talen kommer ifrån, och till höger i tabellen antalet tal studerades från tillhörande distrikt. Benford uppmärksammade att de flesta datamängderna var "lite" BF likt, men framförallt att medelvärdet av unionen följer BF. Speciellt om man tar hänsyn till de givna osäkerhetsmarginalerna.[Ben]

$P(d_m) = \log_{10}(d_{m+1}) - \log_{10}(d_m), m \in \{1, 2, 3, ..., 9\}$
$d_1 = 0.301, \quad d_2 = 0.176, \quad d_3 = 0.125,$
$d_4 = 0.097, \quad d_5 = 0.079, \quad d_6 = 0.067,$
$d_7 = 0.058, \quad d_8 = 0.051, \quad d_9 = 0.046$

PERCENTAGE OF TIMES THE NATURAL NUMBERS 1 TO 9 ARE USED AS FIRST DIGITS IN NUMBERS, AS DETERMINED BY 20,229 OBSERVATIONS

Group	Title	First Digit									Count
		1	2	3	4	5	6	7	8	9	
A	Rivers, Area	31.0	16.4	10.7	11.3	7.2	8.6	5.5	4.2	5.1	335
B	Population	33.9	20.4	14.2	8.1	7.2	6.2	4.1	3.7	2.2	3259
C	Constants	41.3	14.4	4.8	8.6	10.6	5.8	1.0	2.9	10.6	104
D	Newspapers	30.0	18.0	12.0	10.0	8.0	6.0	6.0	5.0	5.0	100
E	Spec. Heat	24.0	18.4	16.2	14.6	10.6	4.1	3.2	4.8	4.1	1389
F	Pressure	29.6	18.3	12.8	9.8	8.3	6.4	5.7	4.4	4.7	703
G	H.P. Lost	30.0	18.4	11.9	10.8	8.1	7.0	5.1	5.1	3.6	690
H	Mol. Wgt.	26.7	25.2	15.4	10.8	6.7	5.1	4.1	2.8	3.2	1800
I	Drainage	27.1	23.9	13.8	12.6	8.2	5.0	5.0	2.5	1.9	159
J	Atomic Wgt.	47.2	18.7	5.5	4.4	6.6	4.4	3.3	4.4	5.5	91
K	$n^{-1}, \sqrt{n}, \cdots$	25.7	20.3	9.7	6.8	6.6	6.8	7.2	8.0	8.9	5000
L	Design	26.8	14.8	14.3	7.5	8.3	8.4	7.0	7.3	5.6	560
M	*Digest*	33.4	18.5	12.4	7.5	7.1	6.5	5.5	4.9	4.2	308
N	Cost Data	32.4	18.8	10.1	10.1	9.8	5.5	4.7	5.5	3.1	741
O	X-Ray Volts	27.9	17.5	14.4	9.0	8.1	7.4	5.1	5.8	4.8	707
P	Am. League	32.7	17.6	12.6	9.8	7.4	6.4	4.9	5.6	3.0	1458
Q	Black Body	31.0	17.3	14.1	8.7	6.6	7.0	5.2	4.7	5.4	1165
R	Addresses	28.9	19.2	12.6	8.8	8.5	6.4	5.6	5.0	5.0	342
S	$n^1, n^2 \cdots n!$	25.3	16.0	12.0	10.0	8.5	8.8	6.8	7.1	5.5	900
T	Death Rate	27.0	18.6	15.7	9.4	6.7	6.5	7.2	4.8	4.1	418
Average.......		30.6	18.5	12.4	9.4	8.0	6.4	5.1	4.9	4.7	1011
Probable Error		±0.8	±0.4	±0.4	±0.3	±0.2	±0.2	±0.2	±0.2	±0.3	—

Newcombs observation föll i glömska eftersom man för tillfället inte hade något användningsområde/intresse för denna typ av "observation". Först 57år senare uppmärksammas lagen igen av fysikern Frank Benford. Som genom att studera data från över oberoende 20 domäner innehållande över 20.000 tal av varierande storleksordningar, kunde Benford komma fram till fördelningen

$$P(d) = \log_{10}(\frac{d+1}{d}).$$

Efter Benford studerat sin data publicerade Benford artikeln *The Law of Anomalous Numbers April 22 år 1938*, där Benford gav förklaringarna till varför ledande siffror skulle fördela sig enligt BF. Dock publicerade Benford aldrig något bevis, utan nöjde sig visa lagen experimentellt. Notera att Benford visade att lagen gäller för unionen av oberoende datamängder. Teodore P.Hill lyckade bevisa för specifika grupper år 1995[Hi1].

5 Kopplingen mellan logaritmiska skalan och Benford's Lag

5.1 Likformiga sannolikhetsfördelningen

Likformiga sannolikhetsfördelningen det unika utfallsrummet där alla utfall har samma sannolikhet.

Notationen $\mathbb{U}[0,1)$ menas att alla slumpvariabler $X \in \mathbb{R}$ mellan $[0,1)$ har samma utfallssannolikhet. Givet ett $0 \leq a < b < 1$ blir utfalls sannolikheten för intervallet $P([a,b]) = b - a$. När $X \sim \mathbb{U}[0,1)$ används, menas approximeras vara likformig inom intervallet $[0,1)$.

5.2 logaritmiska skalan

En logaritmisk skala är en icke linjär skala och används huvudsakligen till att sammaställa/illustrera data som sträcker sig över flera storleksordningar, exempelvis richterskalan. Genom att välja ut en storhet som logaritmen skall förhålla sig till, skalas den utavalda datan till potenser av storheten. Den logaritmiska skalan används blandannat till att omvandla snabbt exponentiellt växande funktioner till mer övergripliga linjära funktioner.

I dagens moderna samhälle använder oss av ett positionssystem med en talbas av 10. Så den tiologaritmiska skalan är på så sätt naturligt förknippat med vårat talsystem. Tiologaritmen är en viktig del i förståelsen av BF.
Där $[x,y) \mapsto_{10}$ *menas* $[x,y) \mapsto [\log_{10}(x) \pmod 1, \log_{10}(y) \pmod 1))$. Observera att vi använder oss av modulo 1 eftersom vi intresserar oss av decimaldelen.
Studerar vi intervallen $[1,2) \mapsto_{10} [0, 0.301)]$ samt intervallet $[2,3] \mapsto_{10} [0.301, 0.477]$. Uppmärksammar vi att längden av intervallen är exakt samma som Benford's sannolikhetsfördelning.

$$||\mathscr{B}_m|| = P_{BF}(d_m) = \log_{10}(d_{m+1}) - \log_{10}(d_m), \forall m \in \{1,2,..,9\}$$

Benford intervall $\mathscr{B}_m$ är intervallen efter transformationen

$$[d_m, d_{m-1}) \mapsto_{10} \mathscr{B}_m \qquad m \in \{1,2,..,9\}$$

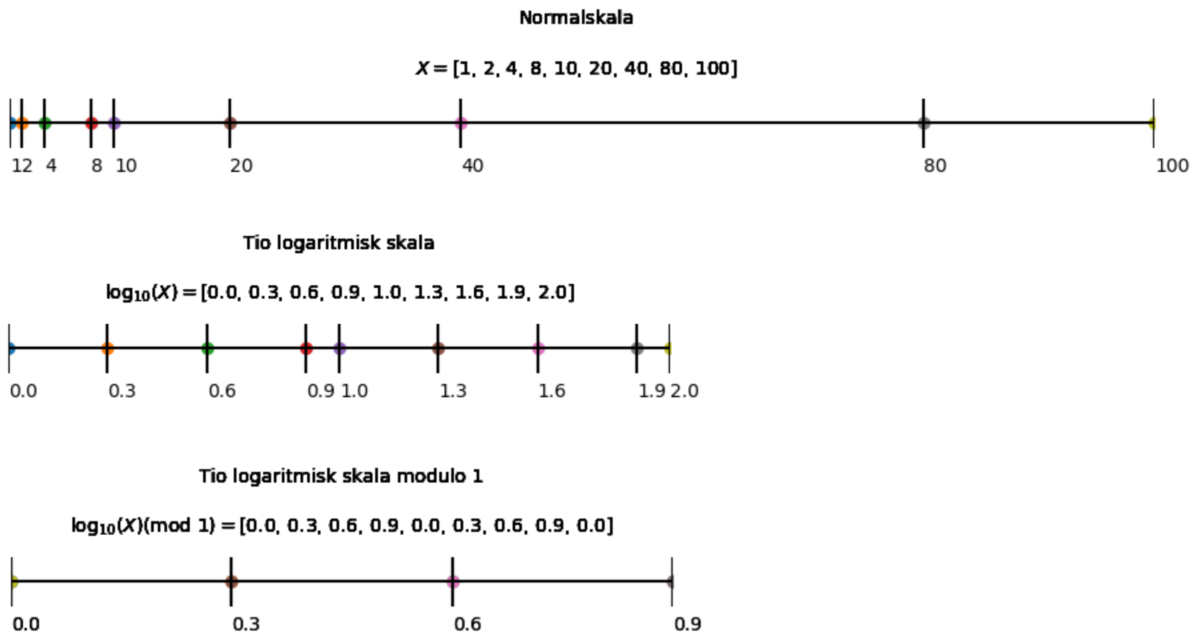
$$\mathscr{B}_m = [\log_{10}(d_m), \log_{10}(d_{m+1})), \qquad m \in \{1,2,..,9\}.$$

Exempel $\mathscr{B}_1 = [0, 0.301), \mathscr{B}_2 = [0.301, 0.477)$.

Definition 1. *Mantissafunktionen* $\mathcal{M}(x)$ av ett godtyckligt tal $x \in \mathbb{R}^+$ är

$$x = \mathcal{M}(x) * 10^n, \qquad n \in \mathbb{Z}, 1 \leq \mathcal{M}(x) < 10$$

Visualisering av tranformationen $X \mapsto \log_{10} X \mapsto \log_{10} X \pmod 1$



Exempel:
Givet en mängd innehållande två godtyckliga strikt positiva tal $\Omega = \{x_1, x_2\}$. Vi utför transformationen

$$x \mapsto \log_{10} x \pmod 1 \qquad \forall x_i \in \Omega$$

Om båda talens decimaldel av mantissan $\mathcal{M}(x) \pmod 1$ efter transformationen ligger inom samma BF-intervall $\mathscr{B}_m$

$$\log_{10}(\frac{1+d_m}{d_m}) \leq \mathcal{M}(x) \pmod 1 < \log_{10}(\frac{1+d_{m+1}}{d_{m+1}}), \qquad m \in \{1, 2, 3, .., 9\},$$

hade dessa tal samma ledande siffra innan transformationen.

Sats 2. Låt X vara en slumpvariabel. För att förenkla antar vi att $X > 0$. Om $\log_{10} X \pmod 1$ är likformigt fördelad följer X Benfordsfördelning.

Bevis. Låt $d_m(X)$ vara den ledande siffran av X . Då gäller

$$d_m(X) = k, \qquad k \in \{1, 2, .., 9\}$$

$$\begin{aligned} P(d_m(X) = k) &= P(X \in \textstyle\bigcup_{n=-\infty}^{\infty} [k, k+1) \times 10^n, \qquad n \in \mathbb{Z} \\ &= P(\log X \in \textstyle\bigcup_{n=-\infty}^{\infty} [\log_{10}(k), \log_{10}(k+1)) + n \\ &= P(\log X \pmod 1 \in [\log_{10}(k), \log_{10}(k+1))) \\ &= \log_{10}(k+1) - \log_{10} k \end{aligned}$$

Definition 3. Givet $a < b < a + 1$, anta att x är ett tal som uppfyller $a \leq x < b \pmod 1$. Om det existerar ett heltal n så att $a < x + n < b$. Om talet n existerar är talet *unikt.*

Sats 4. Givet en slumpvariabel $X \in \mathbb{R}^+$ som uppfyller $\log_{10} X \pmod 1$ är likformigt fördelat. Samt en godtycklig slumpvariabel $Y \in \mathbb{R}^+$. Då följer $X \cdot Y$ Benfordsfördelning.

Bevis. Låt $U = \log_{10} X$ och $V = \log_{10} Y$. Vi antar att $U \pmod 1$ är likformigt fördelad. Samt att V är godtycklig. Låt det existera $0 \leq a < b < 1$ så att

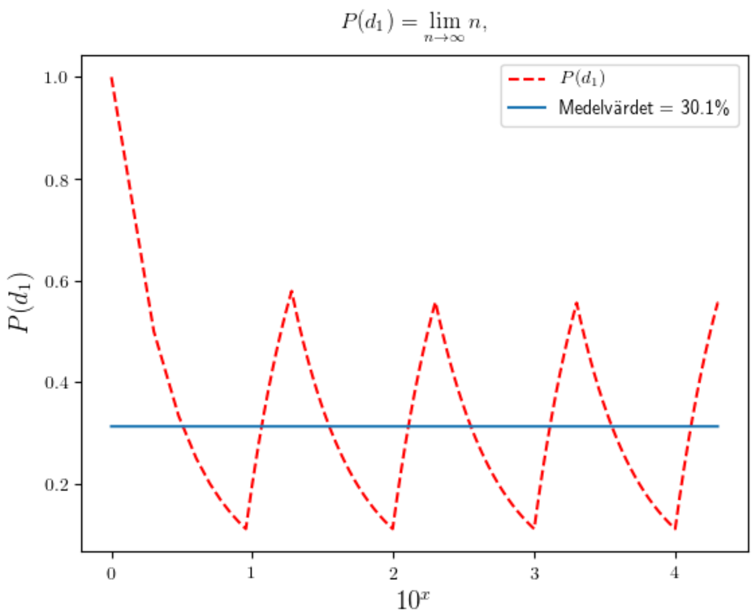
$$\begin{aligned} P(a \leq U + V \pmod 1 < b) &= P((a - V \leq U < b - V) \pmod 1)) \\ &= P(a < U < b), \textit{ eftersom } U \textit{ är likformigt fördelad följer} \\ &= b - a \end{aligned}$$

Vilket bevisar att $U+V \pmod 1$ är likformigt fördelad och därav följer $X \cdot Y$ Benfordsfördelning.

6 Förklarningar till Benford's lag

6.1 Den klassiska bevis metoden

Benford samt många andra matematiker innan Ted P.Hills bevisning, försökte bevisa lagen genom observera att all tal i intervallen $[1, 2), [10, 20), [100, 200]...[10^n, 2 * 10^n)$ kommer att ha d_1 som ledande siffra. Genom att bevisa gränsvärdet $P(d_{m=1})$ när $\lim_{n\to\infty} n$ skulle gå mot Benford sannolikheten $P_{BF}(d_1)$.



Notera mönstret, vilket kommer återupprepas oändligt många gånger när $\lim_{n\to\infty}$.

Men eftersom gränsvärdet saknas så det går inte att bevisa Benford's lag detta sätt.

Många skrifter om Benford's lag samt online föreläsare hävdar att medelvärdet av gränsvärdet kommer vara det eftertraktade värdet 30.1 % . Vilket är en sanning med modifikation.

Anledningen till varför många hävdar att medelvärdet av gränsvärdet är 30.1%, det beror de att samplat lika många mätpunkter mellan $[1, 10)$ om mellan $[10, 100)$. Vilket i sin tur kommer leda till att avståndet mellan mätpunketerna kommer öka som en exponentiellt växande funktion(se (6.2) Gemometriska förklaringen). Annars om vi skulle väljer en slutpunkt x för att kunna illustrera $P(d_1) \lim_{n\to x} n$ blir medelvärdet av gränsvärdet helt beroeende av slutpunkten. Observera att ovanståeende figur har en vald endpunkt för att illustrera det eftertraktade $P_{BF}(d_1)$.

6.2 Geometriska förklaringen

Om man studerar en exponentiellt växande process, exempelvis placerar en svensk krona(k) i en aktie som konstant växer med 3% varje år. Beräkningen $(k+1.03)^{\text{år}}$ ger oss vad aktien är värde efter visst antal år. Tiden det tar för aktien att dubbleras i värde blir således

$$k * 1.03^{\text{år}} = (k+1) \Leftrightarrow \text{år} = \log_{10}\left(\frac{(k+1)}{k} \times \frac{1}{1.03}\right) = \frac{\log_{10}\frac{k+1}{k}}{\log_{10} 1.03}.$$

Tiden det tar för att våran aktie öka från en till två kronor är, dvs att $k_0 \mapsto k_0 + 1$.

$$k_1 = \frac{\log_{10}\frac{k_0+1}{k_0}}{\log_{10} 1.03} = \frac{\log_{10}\frac{2}{1}}{\log_{10} 1.03} = 23.4497\text{år}.$$

Jämtför med tiden det tar för våran aktie att öka i värde från två till tre kronor, $k_1 \mapsto k_1+1$

$$k_2 = \frac{\log_{10}\frac{3}{2}}{\log_{10} 1.03} = 13.7172\text{år}.$$

Tiden det tar att öka från en till tio kronor blir således

$$k_{sum} = \sum_{k=1}^{10} \frac{\log_{10}\frac{k+1}{k}}{\log_{10} 1.03} = 77.8985\text{år}.$$

Om vi nu studerar vi andelen tid kronan spenderade för att öka från en till två, två till tre kronor, etc. Får vi följande

$$d_1 = \frac{k_1}{k_{sum}} = 0.301, \qquad d_2 = \frac{k_2}{k_{sum}} = 0.176, \qquad ..., \qquad d_9 = \frac{k_9}{k_{sum}} = 0.046,$$

$$\frac{\frac{\log_{10}\frac{d+1}{d}}{\log_{10} 1.03}}{\frac{\log_{10} 10}{\log_{10} 1.03}} = \log_{10}(\frac{d+1}{d}) = P_{BF}(d_m).$$

Vilket är exakt BF, notera att beräkningen är generell, dvs ej beroeende på våran ökning av just 3%. Detta resulterar i att samtliga exponentiellt växande serier och funktioner kommer föja BF, likaså alla exponentiellt avtagande serier och funktioner (exempelvis halveringstid hos radioaktiva ämnen).[Mi]

Detta är anleningen varför fibonaccis talföljd, antalet människor i vaje unikt land, aktier etc följer BF.

6.3 Slumpen av slumpade tal

Hjälper Benford's lag dig att vinna på lotto? Svar nej! Eftersom lottodragning är det slumpen som avgör, för i lotterier har alla lotter samma sannolikhet att bli dragen.
Däremot om vi slumpar fram talen $n_1, n_2, n_3...n_i$ och sedan multiplicerar talen med varandra $N_1 = (n_1 \times n_2 \times ... \times n_i)$ vilket bildar det "*genererade slumptalet*" talet N_1. Därefter repeterar denna procedur väldigt många gånger får vi mängden $\Omega = \{N_1, N_2, N_3, ..., N_k\}$. När vi låter $\lim_{k\to\infty}$ kommer våran mängd Ω med stor sannolikhet vara BF.

Låt oss nu göra en tärningsgenerator, som kastar 20 tärningar och multiplicerar prickarna som är uppåt. Då kommer du generera tal melllan $[1, 6^{20}]$, gör vi ovanståeende procedur 9999 gånger då kommer du med högsta sannolikhet en mängd tal som följer BF.

```
Andelen element vars ledande siffra är en 1 är 30.116% Benford är approx 30.1%
Andelen element vars ledande siffra är en 2 är 17.604% Benford är approx 17.6%
Andelen element vars ledande siffra är en 3 är 12.432% Benford är approx 12.5%
Andelen element vars ledande siffra är en 4 är 9.702% Benford är approx 9.7 %
Andelen element vars ledande siffra är en 5 är 8.082% Benford är approx 7.9 %
Andelen element vars ledande siffra är en 7 är 5.701% Benford är approx 6.7 %
Andelen element vars ledande siffra är en 8 är 5.531% Benford är approx 5.8 %
Andelen element vars ledande siffra är en 9 är 4.021% Benford är approx 4.6 %
```

Output från ett pythonsqript som faktoriserar 20 tärningskast ($N_1 = \{n_1 \times n_2, ..., n_{20}\}$), sedan upprepar detta 9999 gånger ($\Omega = \{N_1, N_2, ..., N_{9999}\}$) här visas sannnolikeheten för att varje element ledande siffra.

Eftersom vi multiplicerar flertalet oberoeende variabler kommer enligt centrala gränsvärdessatsen våran mängd $\Omega = \{N_1, N_2, N_3...N_i\}$ vara approximativt normalfördelad.
Viktigt att notera är att endast en del av alla normalfördelningar kommer följa Benfords, men däremot kommer med stor sannolikhet många sammaslagna oberooende normalfördelningar att följa Benfords lag[Mi].
Ett exempel på en normalfördelning som ej är Benfordfördelad är människors längd. Eftersom nästa alla människor är mellan 1-2 meter långa samt att ingen längre än tre meter. Vilket är en väldigt tydligt inte BF.

7 Svagt Benfordfördelad

Svag benfordfördelning menas

$$P_{bf}(d_m) \approx \log_{10}(d_{m+1}) - \log_{10} d_m, \qquad m \in \{1, 2, .., 9\}.$$

Varför vi uppmärksammar svag benfordfördelning(P_{bf}) beror på att datamängder som följer P_{bf} tenderar att följa P_{BF}. Samt att den svaga benfordfördelningen förklarar bättre vilka datamängder som i regel följer Benfordfördelningen. (Notera skillnad mellan stor och liten bokstav)

7.1 Centrala gränsvärdessatsen

Den centrala gränsvärdessatsen är en fundamental sats inom statistik. Enligt centrala gränsvärdessatsen gäller att om man adderar ett stort antal oberoende slumpmässiga variabler med ändliga varianser, kommer summan att gå mot en normalfördelning.[Wiki2]

7.2 Naturlig data

Naturlig data är en mängd tal som är insamlad med samma metod alternativt ur samman distrikt, vars tal ej har manipulerats, ej är stickt bundna och inte helt slumpmässiga. Naturlig data har ofta en tendens till att följa BF och det beror på att datan ofta påverkars naturliga händelseförlopp samt av små slumpmässiga händelser eller incidenter. Dessa slumpvariabler gör att den naturliga datan tenderar att följa centrala gränsvärdessatsen. Samt vara svagt benfordfördelad.

Statistikern William Feller försökte bevisa korrelationen mellan BF och normalfördelningen, genom att hävda att alla normalfördelade mängder som uppfyller

(i) *En possitiv slumpvariabel X följer BF om och endast om* $\log_{10} X \pmod 1$ *är* $\mathbb{U}[0,1)$.

(ii) *Om spridningen av en slumpvariabel X är väldigt stor, då kommer* $\log_{10} X \pmod 1 \sim \mathbb{U}[0,1)$.

(iii) *Om* $\log_{10}$ *har en väldigt stor spridning då kommer* $\log_{10} X \pmod 1 \sim \mathbb{U}[0,1)$.

Tenderar att följa BF. Observera att alla mängder som uppfyller (i), kommer följa Benford lag exakt. Medans de övriga förklara vilka mängder som troligtvis är svagt benfordfördelad.

7.3 Spridnings hypotesen

Att förstå vilka datamängder som bör samt inte bör följa Benford's lag är knepigt att bevisa rent matematiskt. Men det finns några tummregler för vilka datamänder troligtvis kommer följa BF. Observera att en innebörden av en tummregel är att regeln oftast är sann, men är ej en absolut sanning. Dvs alla mängder som uppfyller tummreglerna följer inte Benford's lag. Med hjälp av Feller kan vi skapa tummreglerna

$$
\begin{array}{ll}
(i) & \textit{Data mängder som utbreder sig över många storleksordningar.} \\
(ii) & \textit{Datan visualiserat i en graf har en skevhet åt höger.} \\
(iii) & \textit{Datan får ej vara strikt bunden.} \\
(iv) & \textit{Data där variansen är stor.}
\end{array}
$$

Dessa tummregler ger en förklaring till varför ihopsatta datamängder från olika fördelningar har en tendens att följa Benford's lag (*exempel Benford's datas "averages" se p.10*). Notera att endast "averages" är Benfordfördelad medans datan ur ditrikten tenderar att vara benfordfördelad.

Exempel:
Låt två oberoeende mängder $X = \{x_1, x_2, x_3...x_i\}), Y = \{y_1, y_2, y_3....y_j\}$ där $\forall x, y \in \mathbb{R}^+$. Där Ω är unionen av mänderna enligt

$$\Omega = X \cup Y = \{x_1, x_2, x_3...x_n\} \cup \{y_1, y_2, y_3....y_i\} = \{x_1, x_2, x_3...x_n, y_1, y_2, y_3....y_j\}.$$

Om vi antar att variansen av $Var(\Omega) > Var(Y) \geq Var(X)$. Vilket i sin tur gör att mängden Ω troligtvis kommer att flatna, stäcka sig över fler storlekordningar, vilket restulterar i en bedare och mjukare sanolikhetskurva. Som i sin tur ger en mer likformid fördelning $\log_{10} \Omega \pmod 1$. vilket leder till att chansen för mängden Ω följer bf är troligtvis större än chansen för X och Y separat. För ett mer matematisk bevis se [BH4 p.100-120], pga på bevisets svårighetsgrad utelämnas bevisningen.

8 Introduktion av Borel sigma algebra

Denna uppsats inte handlar om ämnet sigma algebra därför kommer vi endast förklara de delar som berör Benfordslag. Vi kommer använda oss av sigma algebra eftersom denne utesluter mängder som är överdrivet komplicerade, samt mängder inte är relevanta för diskussionen alternativt bevisningen.

8.1 Sigma algebra

Sigma algebran är av stor betydelse inom måtteori samt sannolikhetsteori, där sigma algebran innehåller alla mätbara mängder. En sigma algebra $\mathscr{A}$ av en mängd $\sigma(\Omega)$ är en familj delmängder av Ω som uppfyller följande:

(i) *Tomma mängden är i sigma algebran(* $\emptyset \in \mathscr{A}$*).*
(ii) *Sluten under Komplement* $(A \in \mathscr{A} \Rightarrow A^{\mathscr{K}} = \{\omega \in \Omega : \omega \notin A\} \in \mathscr{A})$.
(iii) *Sluten under uppränkliga unioner* $(A_n \in A$ *för alla* $n \in \mathbb{Z} \Rightarrow \bigcup_{n=\mathbb{Z}} A_n \in \mathscr{A})[BH4]$.

8.2 Borel sigma algebra

En Borelmängd är en mängd som är genererad av öppna mängder. En Borelmängd är en uppräknelig union av öppna mängder och komplement till öppna mängder. Alla Borelmängder är element i den sigma-algebra som genereras av de öppna mängderna, vilken kallas Borel sigma algebra. [Wiki3]

Så om vi kallar intervallet $[1, 10) = \Omega_1$ och dess öppna delintervall $\mathcal{C}_0$. Kallas snittet av $\sigma(\mathcal{C}_0)$ för Borel σ-algebran för $[1, 10)$,och noteras $\mathcal{B}[1, 10)$. Elementen i $\mathcal{B}[1, 10)$ kallar vi för Borelmängder.

9 Mantissa algebra

9.1 Mantissa

Matissan $\mathcal{M}(x)$[Def 1 p.12] har två viktiga egenskaper

$$\begin{aligned} &(i) \quad \mathcal{M}(x) = \mathcal{M}(x \times 10^n), \qquad n \in \mathbb{Z}. \\ &(ii) \quad \mathcal{M}(\mathcal{M}(x)) = \mathcal{M}(x). \end{aligned}$$

Definition 5. Mantissa algebran $\mathscr{M}$ är sigma algebran på $\mathbb{R}^+$ genererad av mantissa funktionen $\mathcal{M}$.

$$\mathscr{M} = \mathbb{R}^+ \cap \sigma(\mathcal{M})$$

Där av följer

$$A \in \mathscr{M} \qquad \Leftrightarrow \qquad A = \bigcup_{n\in\mathbb{Z}} B \times 10^n, \qquad \text{där } B \subseteq \mathcal{B}[1, 10).$$

För att kunna bevisa att Benford's lag är skalinvariant. Måste vi använda rätt utfallsrum, vilket är ett utfallsrum inom matissa algebran $(\mathbb{R}^+, \mathscr{M})$.

Mantissa algebran har följande två axiom

(i) $\mathscr{M}$ *är självlik* ($A \times 10^n = A$ *för alla* $A \in \mathscr{M}$ *eftersom* $\mathcal{M}(x) = \mathcal{M}(x \times 10^n), n \in \mathbb{Z}$)
(ii) $\mathscr{M}$ *är sluten under multiplicering av en skalär* $(\forall\alpha > 0, A \in \mathscr{M} \Rightarrow \alpha A \in \mathscr{M})$.[Hi2]

Sats 6. Givet en godtycklig konstant $\alpha > 0$, behövs endast fallet $1 < \alpha < 10$ undersökas. Eftersom $A \times 10^n = A$ för alla $n \in \mathbb{Z}_{\neq 0}, A \in \mathscr{M}$.

Bevis. Låt C vara ett intervall, $C = [a, b]$ där $1 \leq a < b < 10$. Multiplicering av en godtycklig konstant $\alpha > 0$ ger oss tre möjliga utfall

$$C = \begin{cases} \alpha b \leq 10, & \alpha C \subset [1, 10], & \alpha A = \cup_{n\in\mathbb{Z}} C \times 10^n \in \mathscr{M} \\ \alpha a \leq 10 < \alpha, & C' = [1, \frac{\alpha b}{10}) \cup [a\alpha, 10) & \alpha A = \cup_{n\in\mathbb{Z}} C' \times 10^n \in \mathscr{M} \\ 10 < \alpha, & C' = [\frac{\alpha a}{10} \frac{\alpha b}{10}) & \alpha A = \cup_{n\in\mathbb{Z}} C' \times 10^n \in \mathscr{M} \end{cases}$$

$$\left[\begin{array}{c} \textit{För alla } \alpha > 0,\ A \in \mathscr{M} \textit{ behöves endast fallet } 1 \leq \alpha < 10 \textit{ undersökas.} \\ \alpha A \times 10^n = \alpha A, A \in \mathscr{M}, n \in \mathbb{Z} \end{array} \right]$$

9.2 Benford i utfallsrummet $(\mathbb{R}^+, \mathscr{M})$

När vi befinner oss i ufallsrummet $(\mathbb{R}^+, \mathscr{M})$ kan vi skapa en kontinuelig $P_{\mathbb{BF}}$ vilket en mer *idealiska Benfordfördelning.* Vilken är matematisk starkare än den vanliga P_{BF} eftersom denne är diskriet för $d_m \in \{1, 2, .., 9\}$. Vilket medför att allt som är sannt gällande $P_{\mathbb{PF}}$, är sant för den vanliga P_{BF}.

$$P_{\mathbb{BF}}\left(X \in \bigcup_{n\in Z} [a, b] \times 10^n\right) = \log_{10} b - \log_{10} a, \qquad 1 \leq a < b < 10,$$

$$\Rightarrow P_{BF}(d_m) = \log_{10}(d_{m+1}) - \log_{10}(d_m).$$

Exempel:
Givet en strikt positiv slumpvariabel X som följer Benfordsfördelning, då ges sannolikheten

för att ledande siffran är tre av följande

$$P_{BF}(d_{m=3}) = P_{\mathbb{BF}}\left(X \in \bigcup_{n\in Z}[3,4]\times 10^n\right) = \log_{10} 4 - \log_{10} 3.$$

Sats 7. Benford's lag är skalinvariant

$$P_{\mathbb{BF}}(A) = P_{\mathbb{BF}}(\alpha A), \qquad \forall \alpha > 0, A \in \mathscr{M}.$$

Bevis. Låt det existera ett $A \in \mathscr{M}$

$$A = \bigcup_{n\in\mathbb{Z}}[a,b]\times 10^n, \qquad 1 \leq a < b < 10.$$

Låt $\alpha > 0$. Vi vill visa att $P_{\mathbb{BF}}(A) = P_{\mathbb{BF}}(\alpha A)$ vi kan anta att $1 \leq \alpha < 10$. Där efter följer

$$\alpha A = \bigcup_{n\in\mathbb{Z}}[a,b]\times \alpha 10^n = \bigcup_{n\in\mathbb{Z}} C \times 10^n.$$

Vilket ger oss tre möjliga utfall av C

$$C = \begin{cases} [a\alpha, b\alpha] & , b\alpha \leq 10 \\ [a\alpha, b\alpha) \cup [1, \frac{b\alpha}{10}) & , a\alpha < 10 < b\alpha \\ [\frac{1\alpha}{10}, \frac{b\alpha}{10}] & , 10 \leq a\alpha. \end{cases}$$

Vilket gör att sannolikheten för αA blir således

$$P_{\mathbb{BF}}(\alpha A) = \begin{cases} \log_{10}(\frac{b}{\alpha}) - \log_{10}(\frac{a}{\alpha}) \\ 1 - \log_{10}(\frac{b}{\alpha}) - \log_{10}(\frac{a}{\alpha}) - 1 & \textit{där alla} = \log b - \log a \\ \log_{10}(\frac{b\alpha}{10}) - \log_{10}(\frac{a\alpha}{10}) \end{cases}$$

Vilket bevisar att Benfordsfördelning är skalinvariant.

10 Mer intressanta egenskaper

10.1 Basinvarians

Benford's har även en bas invariant egenskap, betyder att mängder som följer BF, kan vi utföra ett basbyte på elementen. Och efter basbytet kommer mängden forsatt följa BF. Pga bevisets svårighetsgrad utelämnas beviset i detta arbete. Benford's lag för en godtycklig bas blir således

$$P(d) = \log_b(d+1) - \log_b(d) = \log_b\left(1 + \frac{1}{d}\right), \qquad b \leq 2, d \in \{1, 2, 3, .., 9\}.$$

där b är basen. [Hi1]

10.2 Förutsäger mer än den första siffran

Det finns även den generella BF som är den används för att beräkna sannolikheten för en specifik sifferkombination.

$$P(D_1, D_2, D_3...D_m) = \log_{10}\left(1 + \left(\sum_{j=1}^{m} 10^{m-j} d_j\right)^{-1}\right)$$

Exempel är sannolikheten för att $P(3, 14) = P(d_3, D_1, 4) = \log_{10} \frac{315}{314} = 0.1138$ (Denna lag bevisas/förklaras analogt med Benford's lag(en ledande siffra), samt är också skal och basinvarant[BH4])

10.3 Upptäcka bedrägeri med hjälp av Benfords'lag

Eftersom naturlig data har en tendens att vilja följa BF, insåg ekonomen Hal Varian(1972) även privatpersoners/företags skattedeklarationer borde ungefärligt följa BF. Vilket Mark Nigrini[Wiki2] visade stämmde hyfsat bra. Eftersom miljoners människkor skattedeklarerar vaje år krävs det att undersöknings programvaran som letar efter skattefusk, skall kunna hantera stora datamängder snabbt och effektivt.
Eftersom Benford's lag endast kärver att en dator kontrollerar den/de första siffran/sifforna ur varje tal. Kan datorer effektivt BF granska stora mängder data. Deklarationer om visade sig groft bryta mot BF sållades ut för att senare kontrollera med andra metoder, som sin tur kan bekräfta fusk.

Orsaken till att deklarationer inte följer BF exakt beror till stor del, att produkter som vi inhandlar har en manipulerad prissättning, exempel produkter säljs ofta för 99kr istället för 100kr. Samt att vi inte lika frekvent inhandlar varor som överskrider flera storleksordningar. Trots detta är BF en bra indikator på att något är suspekt. Eftersom

BF ger utslag på bedragare som valt ut nummer på måfå alt försöker har ett någorlunda jämnt antal ettor, tvåor osv.

Benford's lag används även för att kontrollera forskarrapporter,eftersom insamlad data(naturlig data) tenderar att följa BF. Vilket gör att en BF kontroll kan misstänka rapporter innehållande felaktig data eller fuskat resultat genom manipulering av siffror av siffor. Benford's lag användes som bevis för valfusk i Iran 2009[Wiki2].

11 Vinnaren av leken

Benford's lag är det vinnande konceptet, om spelare A väljer en mängd tal som följer Benfordslag kommer mängden tal vara skalinvarianta och således kommer sannolikheten för att ledande siffra efter multiplikation men en godtycklig konstant vara

$$P(d_1) + P(d_2) + P(d_3) = \sum_{d=1}^{3} \log_{10}(1 + \frac{1}{d}) = 0.602.$$

Så 60% av alla rundor kommer spelare A att vinna ifall hen använder sig av Benford's lag.

12 Avslutning

Vi har studerat grunderna av Benford's lag. Från den första enkla observationen av nötta logaritmtabeller, till att lagen återupptäcktes experimentellt 57 år senare. En observation och ett experiment som utelämnade matematiska bevis.

Till den relativt enkla lagen visade sig inte finnas en simpel förklaring. Än så länge får vi nöja oss med flera små förklaringar.
Eftersom så många olika datamängder följer lagen, borde det finnas ett globalt samband som förhoppningsvis kommer förklara allt gällande Benford's lag.

Arbetets huvudändamål är att förklara varför naturliga tal följer Benfordslag samt matematiskt bevisa lagens märkligaste egenskap skalinvarians.

13 Källor

[Ben] $Benford, F.$(1938), *The law of anomalous numbers, Proc. Amer. Philosophical Soc. 78, 551–572.*

[$BH3$] *Berger, A. and Hill, T.P. (2011),* *Benford's Law strikes back: No simple explanation in sight for mathematical gem, Math. Intelligencer 33, 85–91.*

[$BH4$] *Berger, A. and Hill, T.P. (2015),* *An Introduction to Benford's Law Princeton University Press, Princeton, NJ, 1-50,100-120.*

[Fel] *Feller, W. (1966),* *An Introduction to Probability Theory and Its Applications vol 2, 2nd ed., J. Wiley, New York.*

[Few] *Fewster, R. (2009),* *A simple Explanation of Benford's Law, Amer. Statist. 63(1), 20–25.*

[$Hi1$] *Hill, T.P. (1995),* *Base-Invariance Implies Benford's Law, Proc. Amer. Math. Soc. 123(3), 887–895.*

[$Hi2$] *Hill, T.P. (1995),* *A Statistical Derivation of the Significant-Digit Law, Statis. Sci. 10(4), 354–363*

[$Hi3$] *Hill, T.p. (1996),* *The first-digit phenomenon, American Scientists, vol. 86,358–363*

[Ne] *Newcomb, S. (1881),* *Note on the frequency of use of the different digits in natural numbers, Amer. J. Math. 9, 201–205*

[Pi] *Pinkham, R. (1961),* *On the Distribution of First Significant Digits, Ann. Math. Statist. 32(4), 1223–1230.*

[Mi] *Miller, S.j. (2015),* *Benford's Law: Theory and Applications Princeton University Press,3-18.*

[Borel] https://www.youtube.com/watch?v=P_ROD0l0hfk
2019-08-12

[Ja] Jamain, A. (2001),
imperial college of london department of mathematics
http://wwwf.imperial.ac.u/~nadams/classificationgroup/Benfords-Law.pdf
2019-08-12

[Wa] https://plus.maths.org/content/looking-out-number-one
2019-08-12

[Wiki1] https://en.wikipedia.org/wiki/Benford%27s_law
2019-08-12

[Wiki2] https://sv.wikipedia.org/wiki/Centrala_gr%C3%A4nsv%C3%A4rdessatsen
2019-08-12

[Wiki3] https://sv.wikipedia.org/wiki/Borelm%C3%A4ngd
2019-08-12

[Wol] http://mathworld.wolfram.com/BenfordsLaw.html
2019-08-12