The Newcomb-Benford Law: Theory and Applications

Christoph Richard

Universität Erlangen

Bielefeld, 25.3.2010

www.mi.uni-erlangen.de/∼richard

Simon Newcomb 1881

logarithm tables: only first pages worn out heavily (digit 1)

The law of probability of the occurrence of numbers is such that all mantissæ of their logarithms are equally probable.

In other words, every part of a table of anti-logarithms is entered with equal frequency. We thus find the required probabilities of occurrence in the case of the first two significant digits of a natural number to be:

Dig.	First Digit.	Second Digit.
0	. . .	0.1197
1	0.3010	0.1139
2	0.1761	0.1088
3	0.1249	0.1043
4	0.0969	0.1003
5	0.0792	0.0967
6	0.0669	0.0934
7	0.0580	0.0904
8	0.0512	0.0876
9	0.0458	0.0850

In the case of the third figure the probability will be nearly the same for each digit, and for the fourth and following ones the difference will be inappreciable.

argument: "natural" random numbers $X > 0$ should obey:

- $\log_{10}(X) \bmod 1$ uniformly distributed on $[0, 1)$
- this implies (frequency first digit k) $= \log_{10}((k+1)/k)$

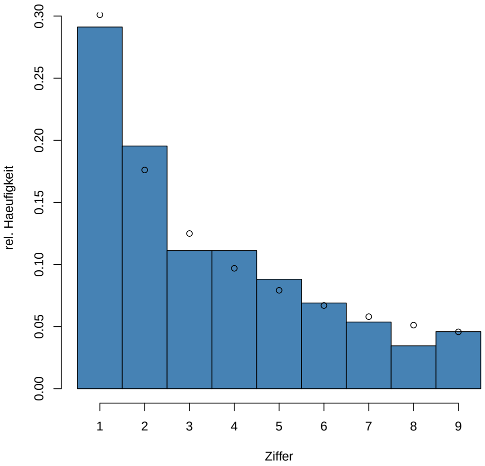
Frank Benford 1938 (independently of Newcomb)

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	—

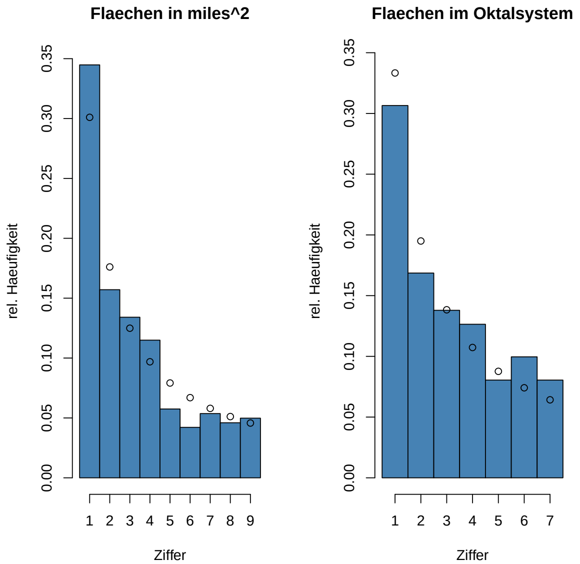
- n^a does not obey NBL (see below)
- data "tuned" through rounding (Diaconis, Freedman 79)
- publication after Bethe et al, The multiple scattering of electrons

example: areas of the 271 states of the world in km^2



(CIA world factbook 2007)

distribution appears to be scale-invariant and base-invariant:



universality of the NBL-digit distribution

random numbers from “natural” phenomena often satisfy:

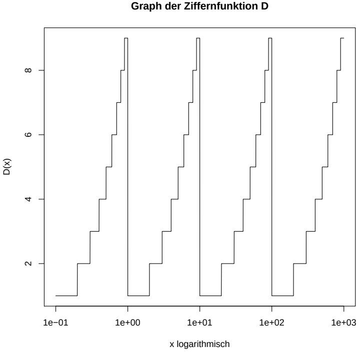
- positive values
- values range over several orders of magnitude
- composed of many (nearly) independent factors
- not artificially processed (rounded, truncated, etc.)

Then typically NBL arises.

mathematical explanation?

digit functions

$D_{10}(x)$ leading digit of $x > 0$ in decimal representation



D_{10} on $(0, \infty)$ Borel-measurable: $D_{10}^{-1}(\{k\}) = \bigcup_{n \in \mathbb{Z}} [k \cdot 10^n, (k+1) \cdot 10^n)$

$D_b(x)$ leading digit of $x > 0$ in base b representation, $b \in \mathbb{N}$, $b > 2$

Definition (Newcomb-Benford Law)

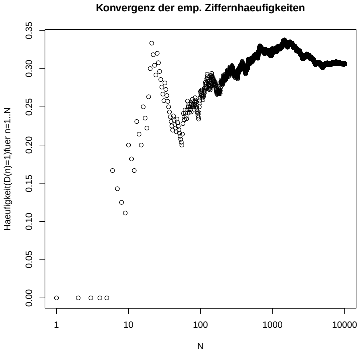
Let $X > 0$ be a positive random variable. X satisfies the Newcomb-Benford law for base b (b-NBL), if

$$\mathbb{P}(D_b(X) = k) = \log_b(1 + 1/k) \qquad (k = 1, \dots, b-1)$$

- Which X satisfy b-NBL?
- For practically every sequence $(x_n)_{n\in\mathbb{N}}$ of b-NBL random numbers the empirical digit frequencies converge to the b-NBL probability.

almost sure convergence of empirical frequencies

$$\frac{1}{n}\mathrm{card}\left(\{1 \le i \le n : D_{10}(x_i) = 1\}\right) \longrightarrow \mathbb{P}(D_{10}(X) = 1) = \log_{10}(2)$$



reason: almost sure convergence of the empirical distribution function (Glivenko-Cantelli theorem, strong law of large numbers) generalisation:

Definition (strong Newcomb-Benford Law)

Let $X > 0$ be a positive random variable.

(i) *X satisfies the strong Newcomb-Benford law for base b (b-sNBL), if for all real numbers $1 \leq \alpha \leq \beta < b$ we have*

$$\mathbb{P}(\exists n \in \mathbb{Z} : \alpha \cdot b^n \leq X < \beta \cdot b^n) = \log_b(\beta/\alpha).$$

(ii) *X satisfies the strong Newcomb-Benford law (sNBL), if X b-sNBL for all $b > 2$.*

- We have b-sNBL $\Rightarrow$ b-NBL. (choose $\alpha := k$, $\beta := k + 1$)
- "natural" b-NBL data appears to be even sNBL.

Theorem

Let $X > 0$ be a positive random variable. Then the following are equivalent:

(i) *X is b-sNBL.*

(ii) *$\log_b(X) \bmod 1$ is uniformly distributed on $[0,1)$.*

reason:

Let $1 \leq \alpha \leq \beta < b$ be arbitrary real numbers. Then

$$\begin{aligned} &\mathbb{P}(\exists n \in \mathbb{Z} : \alpha \cdot b^n \leq X < \beta \cdot b^n) \\ &\qquad = \mathbb{P}\left(\exists n \in \mathbb{Z} : \log_b(\alpha) + n \leq \log_b(X) < \log_b(\beta) + n\right) \\ &\qquad = \mathbb{P}\left(\log_b(X) \bmod 1 \in [\log_b(\alpha), \log_b(\beta))\right) \end{aligned}$$

But this characterises the distribution! □

remarks:

- examples of b-sNBL random variables:

$$X := b^{U+V}, \qquad U \sim \mathrm{unif}[0,1), \quad \mathrm{im}(V) \subseteq \mathbb{Z}$$

  (cf. Leemis et al 2000)
- b_1-sNBL $\not\Rightarrow$ b_2-NBL ($X := 10^U$, $b_1 := 10$, $b_2 := 3$)
- simple example of b-NBL, but not b-sNBL?

Let X b-NBL for $b > 2$. Then also $Y := 2X$ b-NBL:

$$\begin{aligned}\mathbb{P}\left(D_b(Y) = 1\right) &= \mathbb{P}\left(D_b(X) = 2\right) + \mathbb{P}\left(D_b(X) = 3\right) \\ &= \log_b\left(\frac{3}{2}\right) + \log_b\left(\frac{4}{3}\right) = \log_b\left(\frac{2}{1}\right) \\ &= \mathbb{P}\left(D_b(X) = 1\right)\end{aligned}$$

(analogously for other digits)

Definition (b-scale invariance)

A positive random variable $X > 0$ is b-scale invariant, if for all $c > 0$ we have: for all real numbers $1 \leq \alpha \leq \beta < b$ we have

$$\mathbb{P}(\exists n : \alpha \cdot b^n \leq X < \beta \cdot b^n) = \mathbb{P}(\exists n : \alpha \cdot b^n \leq cX < \beta \cdot b^n).$$

Theorem (b-scale invariance I)

Let $X > 0$ be a positive random variable. Then the following are equivalent:

(i) *X is b-scale invariant.*

(ii) *X is b-sNBL.*

reason:

Check characterisation of previous theorem. Let $1 \leq \alpha \leq \beta < b$ be arbitrary real numbers. Then

$$\mathbb{P}(\exists n : \alpha \cdot b^n \leq cX < \beta \cdot b^n) = \mathbb{P}\left(\log_b(cX) \bmod 1 \in [\log_b(\alpha), \log_b(\beta))\right).$$

Use that uniform distribution is the only translation invariant distribution.

□

other digit frequencies

$D_b^{(\ell)}(x)$ ℓ leading digits of $x > 0$ in base b (e.g. $D_{10}^{(3)}(\pi) = (3, 1, 4)$)

Theorem (*b*-scale invariance II)

For a positive random variable $X > 0$ are equivalent:

(i) *X is b-sNBL.*

(ii) *For all $\ell \in \mathbb{N}$ we have: for all choices of $k_1, \ldots, k_\ell$*

$$\mathbb{P}\left(D_b^{(\ell)}(X) = (k_1, \ldots, k_\ell)\right) = \log_b\left(1 + \frac{1}{(k_1 \cdots k_\ell)_b}\right).$$

(iii) *For all $c > 0$ and all $\ell \in \mathbb{N}$ we have: for all choices of $k_1, \ldots, k_\ell$*

$$\mathbb{P}\left(D_b^{(\ell)}(cX) = (k_1, \ldots, k_\ell)\right) = \mathbb{P}\left(D_b^{(\ell)}(X) = (k_1, \ldots, k_\ell)\right).$$

$(k_1 \cdots k_\ell)_b := k_\ell + k_{\ell-1} b + \ldots + k_2 b^{\ell-2} + k_1 b^{\ell-1}$

reason:

(i) $\Rightarrow$ (ii):
choose $\alpha := k_1 b^{1-1} + \ldots + k_\ell b^{1-\ell}$, $\beta := \alpha + b^{1-\ell}$

(ii) $\Rightarrow$ (i):
Approximate α, β by numbers with finitely many digits. Use left continuity in α, β.

(i) $\iff$ (iii): analogously

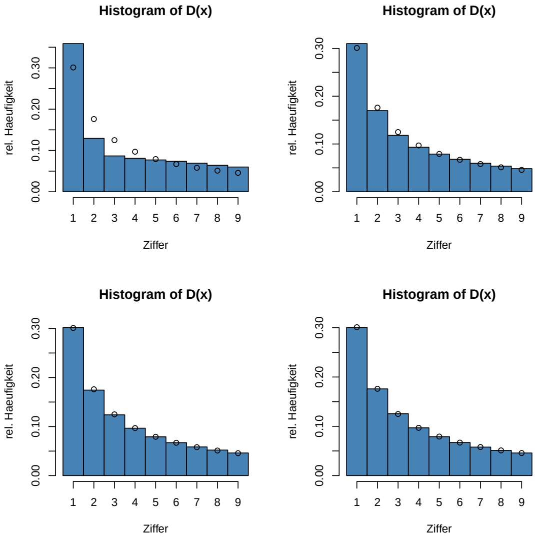
Theorem (universality of sNBL)

Let $X_1, \ldots, X_n$ *independent identically distributed positive random variables with density. We then have for all* $b > 2$*: for all real numbers* $1 \leq \alpha \leq \beta < b$

$$\lim_{n\to\infty} \mathbb{P}\left(\exists n : \alpha \cdot b^n \leq \prod_{i=1}^{n} X_i < \beta \cdot b^n\right) = \log_b(\beta/\alpha).$$

- Hence $\prod_{i=1}^{n} X_i$ approximately satisfies sNBL for n large enough.
- We have approximately scale- and (!) base-invariance.
- Statement remains true for special discrete random variables (see below)
- generalisations (see Miller, Nigrini 2008)

simulation standard normal distribution $n = 1, 2, 3, 4$



proof rests on CLT for S^1-random variables:

Lemma (Lévy 1939)

Let $Y_1, \dots, Y_n$ independent identically distributed S^1-random variables, whose image is not contained in a regular polygon. Then the distribution of $\sum_{i=1}^n Y_i$ converges to the uniform distribution on S^1. □

Universality theorem follows with $Y_i := \log_b(X_i)$.

Newcomb uses essentially the same argument!

almost-NBL

$X = b^Y$ with $\mathbb{V}(Y)$ large should approximately obey b-NBL, e.g.:

Theorem (Dümbgen, Leuenberger 2008)

Let $X = b^Y$ with $Y \sim N(\mu, \sigma^2)$ and $\sigma \geq 1/6$. Let $h(m) = \sqrt{m!/m^m}$. Then we have for all $\ell \in \mathbb{N}$ and for all choices of $k_1, \ldots, k_\ell$

$$\left| \frac{\mathbb{P}\left(D_b^{(\ell)}(X) = (k_1, \ldots, k_\ell)\right)}{\log_b\left(1 + \frac{1}{(k_1 \cdots k_\ell)_b}\right)} - 1 \right| \leq 3h(\lfloor 36\sigma^2 \rfloor).$$

For $\sigma = 1$ already $3h(36) \approx 1.774 \cdot 10^{-7}$!

other example: $X \sim Ex(\lambda)$ (Engel, Leuenberger 2003)

Diaconis 1977: sNBL for sequences $(x_n)_{n\in\mathbb{N}}$ of positive numbers

- equidistribution of $(\log_b(x_n))_{n\in\mathbb{N}}$ modulo 1:

$$\lim_{N\to\infty} \frac{1}{N}\mathrm{card}\left(\{\log_b(x_1) \bmod 1, \dots, \log_b(x_N) \bmod 1\} \cap [a, b]\right) = b - a$$

  for every interval $[a, b] \subseteq [0, 1]$

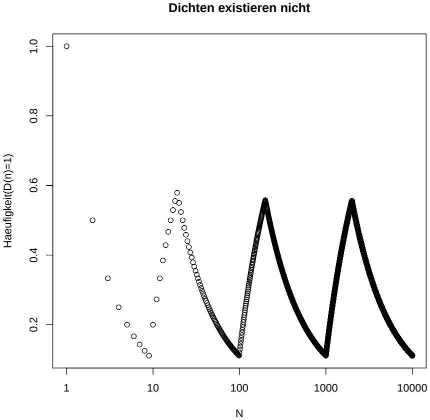
- Above theorems hold *mutatis mutandis*, if probabilities are replaced by limits of empirical densitites, e.g.:

$$\lim_{N\to\infty} \frac{1}{N}\mathrm{card}\left(\{1 \le n \le N : D_b(x_n) = k\}\right) \to \log_b(1+1/k) \quad (N \to \infty)$$

- examples: 2^n, $n!$, Fibonacci numbers
- counter-examples: n^a, $\log_b(n)$

What's wrong with $(n)_{n\in\mathbb{N}}$?

empirical frequencies $\frac{1}{N}\mathrm{card}\left(\{n \in \{1,\dots,N\} : D_{10}(n) = 1\}\right)$



"natural" limit frequency?

sNBL in dynamical systems

sNBL near stable fixpoints, e.g.:

Theorem (Berger, Bunimovic, Hill 2004)

Let $T(x) = \alpha x(1-x)$ *with* $|\alpha| \in (0,1)$. *Then the following are equivalent:*

(i) *Orbit* $(T^{(n)}(x))_{n\in\mathbb{N}}$ *is b-sNBL for all* $x \neq 0$ *sufficiently close to* 0.

(ii) $\log_b |\alpha| \notin \mathbb{Q}$.

- results for non-autonomous systems ($n!$, e^n, Fibonacci, Newton iteration, ...)
- results for differential equations

tests of distribution hypotheses: example "area of countries"

- Newcomb-Benford law (discrete):
  - χ^2-test of NBL with 8 degrees of freedom
  - value of χ^2-statistic: 3.3621, p-value: 0.9096
- strong Newcomb-Benford law (continuous):
  - Kolmogorov-Smirnov-test of $\log(X) \bmod 1$ for uniform distribution
  - 226 largest values (without bindings), two-sided test
  - value of KS-statistic: 0.0343, p-value: 0.9003.

In both examples NBL hypothesis consistent with data.

specific goodness-of-fit tests for NBL (Nigrini 2000, Posch 2005) some examples for applications in economics:

- tax fraud detection
  - book on special methods (Nigrini 2000)
  - being used by tax offices
  - implemented in bookkeeping software (e.g. Audit Commander)
- inflation rate "unbereinigt" vs. "saisonal angepasst" (Posch 2005)
- gross national product of different countries and regimes (Hellan, Nye 2002)

presidential election Iran 2009

- candidates: Ahmadinejad, Mousavi, Karroubi, Rezaei
- analysis of Boudewijn F. Roukema (Torun)
  - data of 366 districts
  - analysis 1. digit: A: 1 too seldom, 2 too frequent; K: 7 too frequent
- Should NBL hold? (number of people per district)
- specialist for NB-analysis of election results: Walter R. Mebane (Michigan)
  - distribution of second digit robust against small deviations of NBL
  - deviations in 12 largest districts for K, R
- deviations to pre-election vote estimates for K,R
- suspicion: votes for K,R transferred to A

more universal distributions

given the above assumptions on a data set, one frequently observes

- Newcomb-Benford law:
  proportions of numbers with first digit k approximately $\log_{10}(1 + 1/k)$.
- Zipf's law:
  n-th smallest value approx. $Cn^{-\alpha}$ for n large ($C > 0$, $\alpha > 0$).
- Pareto distribution:
  proportion of numbers with at least m digits approximately exponentially distributed for large m

(normal distribution, if data concentrated about mean)

some references

- T.Tao, Benford's law, Zipf's law, and the Pareto distribution http://terrytao.wordpress.com/2009/07/03/
- P. Mörters, Benfords Gesetz über die Verteilung der Ziffern (2001)
- online-collection www.benfordonline.net
- T.P. Hill, The first digit phenomenon, American Scientist 86 (1998), 358–364
- M.Nigrini, Digital Analysis Using Benford's Law: Test Statistics for Auditors, Global Audit Publications 2000.
- Iran election 2009 http://election.princeton.edu/2009/06/21/
- N.Hüngerbühler, Benfords Gesetz über führende Ziffern: Wie die Mathematik Steuersündern das Fürchten lehrt, EducETH