Benford' s Law between Number Theory and Probability

Rita Giuliano (Pisa)

Department of Mathematics
University of Pisa
ITALY

Benford's Law for fraud detection
Stresa (IT), July 10-12, 2019

Outline

- Brief history of Benford's Law
- Regular sets and conditional density: an extension of Benford's Law (joint work with Georges Grekos, Université Jean Monnet, St. Etienne)
  - Motivation
  - Results
- A unifying probabilistic interpretation of Benford's Law (joint work with Élise Janvresse, Université de Picardie)
  - Motivation
  - Results

Brief history of Benford's Law

- S. Newcomb (1881): the first pages of logarithmic tables are more consumed than the last ones $\Longrightarrow$ they are used more frequently
- F. Benford (1938; 57 years later!)
  - $\rightarrow$ examinations of data coming from many sources (electricity bills, street addresses...)
  - $\rightarrow$ he rediscovered the same phenomenon.
- nowadays known as **Benford's Law**:
  The "frequency" of the numbers with first significant decimal digit p is
  $$\log_{10} \frac{p+1}{p}$$
- in particular it is not uniform as could be expected!

A number–theoretic formulation

How can we interpret the word "frequency "?

A possible answer

$$A \subseteq \mathbb{N}$$

$$A(x) = \#\big(A \cap [1, x]\big)$$

$=$ number ($\#$) of integers belonging to A and less or equal to x

$\rightarrow$ Attempt of definition of "frequency "of A

$$= \text{"natural" density of } A = d(A) = \lim_{n\to\infty} \frac{A(n)}{n}.$$

$\rightarrow$ Difficulty: For $A_p = \{$integers with first digit $= p\}$ the limit doesn't exist! In fact

$$\underline{d}(A) = \liminf_{n\to\infty} \frac{A(n)}{n} = \frac{1}{9p}; \quad \overline{d}(A) = \limsup_{n\to\infty} \frac{A(n)}{n} = \frac{10}{9p}.$$

A number–theoretic formulation

No density= no frequency?

Let's try to argue more widely. Attach a "weight" $\mu(\{k\}) = 1$ to each integer k. Then

- $$\text{"natural measure" of } (A \cap [1,n]) = \mu(A \cap [1,n])$$
  $$= \sum_{\substack{1 \le k \le n, \\ k \in A}} \mu(\{k\}) = \sum_{\substack{1 \le k \le n, \\ k \in A}} 1 = A(n)$$

- $$\text{"natural measure" of } [1,n] = \mu(\mathbb{N} \cap [1,n]) = \sum_{1 \le k \le n} \mu(\{k\}) = n$$

- $$\frac{A(n)}{n} = \frac{\mu(A \cap [1,n])}{\mu(\mathbb{N} \cap [1,n])}.$$

A number–theoretic formulation

What about other weights? For instance $\mu(\{k\}) = \frac{1}{k}$. Then

- $$\text{“logarithmic measure” of } (A \cap [1,n]) = \mu(A \cap [1,n]) = \sum_{\substack{1 \le k \le n, \\ k \in A}} \frac{1}{k}$$

- $$\text{“logarithmic measure” of } [1,n] = \mu(\mathbb{N} \cap [1,n]) = \sum_{1 \le k \le n} \frac{1}{k}$$

- $$\text{“logarithmic” density of } A = \delta(A) =$$
$$\lim_{n \to \infty} \frac{\mu(A \cap [1,n])}{\mu(\mathbb{N} \cap [1,n])} = \lim_{n \to \infty} \frac{1}{\log n} \sum_{\substack{1 \le k \le n, \\ k \in A}} \frac{1}{k}.$$

The term “logarithmic” comes from

$$\mu(\mathbb{N} \cap [1,n]) = \sum_{1 \le k \le n} \frac{1}{k} \sim \log n.$$

A number–theoretic formulation

$\mathbb{P}$ = set of prime numbers. It is known that, with $\mu(\{k\}) = \frac{1}{k}$

$$\lim_{n\to\infty} \frac{\mu(A_p \cap \mathbb{P} \cap [1,n])}{\mu(\mathbb{P} \cap [1,n])} = \lim_{n\to\infty} \frac{\sum_{1\le k\le n, k\in A_p\cap\mathbb{P}} \frac{1}{k}}{\sum_{1\le k\le n, k\in\mathbb{P}} \frac{1}{k}} = \log_{10} \frac{p+1}{p}.$$

With a term borrowed from probability, we call

$$\lim_{n\to\infty} \frac{\mu(A \cap \mathbb{P} \cap [1,n])}{\mu(\mathbb{P} \cap [1,n])} = \text{logaritmic density of } A, \text{ conditioned to } \mathbb{P}.$$

Questions

So, the conditional logarithmic density of A_p , given $\mathbb{P}$, is equal to its (non-conditional) logarithmic density.

Question 1

Which sets other than $\mathbb{P}$?

Question 2

Which sets other than A_p?

Answer to question 1

Any "regular" set $\mathbb{H}$ will do

What is regularity?

$$(\text{counting function of } \mathbb{H})(x) = H(x) = \#(\mathbb{H} \cap [1, x])$$

$$= \text{number of elements of } \mathbb{H} \text{ that are less or equal to } x$$

Definition

$\mathbb{H} \subseteq \mathbb{N}$ is "regular" with exponent $\lambda \in (0, 1]$ if the function

$$L(x) = \frac{H(x)}{x}$$

is "slowly varying" as $x \to \infty$ i.e.$\sim$ behaves approximately as a constant for large x.

Examples of slowly varying functions: $\log x$, $\frac{1}{\log x}$, $\log \log x$, $\sin \frac{1}{x}$...

Examples of regular sets

- $$\mathbb{H} = \{n^r, n \in \mathbb{N}\} = \text{set of } r\text{–th powers}$$
  $$H(x) = \lfloor x^{\frac{1}{r}} \rfloor \text{is regularly varying with exponent } \lambda = \frac{1}{r}.$$
- $$\mathbb{H} = \text{set of all powers}$$
  $$H(x) \sim \sqrt{x}$$
  $$\Longrightarrow H \text{ is regularly varying with exponent } \lambda = \tfrac{1}{2}.$$
- $$\mathbb{H} = \mathbb{P}$$
  $$(\text{counting function of } \mathbb{P})(x) = \pi(x) \sim \frac{x}{\log x}$$
  $$\Longrightarrow \pi \text{ is regularly varying with exponent } \lambda = 1.$$

Answer to question 2

$$A = \bigcup_n \left([p_n, q_n[\cap \mathbb{N} \right)$$

with

$$p_n \sim \sigma q_n, n \to \infty, \qquad \sigma < 1$$

What about A_p?

$$A_p = \bigcup_n \left([p \cdot 10^n, (p+1) \cdot 10^n[\cap \mathbb{N} \right)$$

(for ex. ($p = 3$): $371 \in [300, 400[= [3 \cdot 10^2, 4 \cdot 10^2[$, so 371 belongs to the second interval ($n = 2$).
In this case

$$p_n = p \cdot 10^n, \qquad q_n = (p+1) \cdot 10^n, \qquad \sigma = \frac{p}{p+1}$$

A probabilistic formulation

Define the

$$\textbf{mantissa in base 10 of } x = \mathcal{M}(x) \in [1, 10[$$

$$\mathcal{M}(x) = 10^{\{\log_{10} x\}}$$

Meaning

$[a]$ = (lower) integer part of a = greatest integer less or equal to a.

$\{a\}$ =fractional part of $a = a - \lfloor a \rfloor$

WARNING!

$\rightarrow \{2, 76\} = 2, 76 - 2 = 0, 76$

$\rightarrow \{-3, 84\} = -3, 84 - (-4) = 0, 16.$

A probabilistic formulation

An example with $x = 0,00487$

$$10^{-3} = 0,001 \leq 0,00487 < 0,01 = 10^{-2}$$
$$\iff -3 \leq \log_{10} 0,00487, -2$$
$$\iff \lfloor \log_{10} 0,00487 \rfloor = -3$$

Using the scientific notation

$$0,00487 = 4,87 \cdot 10^{-3} = 4,87 \cdot 10^{\lfloor \log_{10} 0,00487 \rfloor}$$
$$= 4,87 \cdot 10^{\log_{10} 0,00487 - \{\log_{10} 0,00487\}}$$
$$= 4,87 \cdot 10^{\log_{10} 0,00487} \cdot 10^{-\{\log_{10} 0,00487\}}$$
$$= 4,87 \cdot 0,00487 \cdot 10^{-\{\log_{10} 0,00487\}}$$

A probabilistic formulation

$$\color{red}{4,87} \cdot \underbrace{10^{-\{\log_{10} 0{,}00487\}}}_{=\mathcal{M}(0{,}00487)} = 1$$

$$\Longleftrightarrow$$

$$\mathcal{M}(0,00487) = \color{red}{4,87}$$

i.e.

the mantissa of x is the number which multiplies the power of 10 when x is written in scientific notation.

How to formulate Benford's law in terms of the mantissa ?

The first significant digit of $x = p$

$$\Longleftrightarrow$$

$\mathcal{M}(x)$ **is between** p **and** $p+1$**:**

$$P(\text{the first significant digit of } x = p) = P\big(p \leq \mathcal{M}(x) < p+1\big)$$

Thus Benford's law says that

$$P\big(p \leq \mathcal{M}(x) < p+1\big) = \log_{10} \frac{p+1}{p} = \log_{10}(p+1) - \log_{10} p,$$

or equivalently

For any $1 \leq t \leq 10$**, the proportion of** $x > 0$ **which satisfy** $\mathcal{M}(x) \in [1, t[$ **is**

$$\beta([1, t[) = \log_{10} t$$

Janvresse and De La Rue heuristics:

Consider data as coming from a r.v. on the interval $[0, A]$.

Benford himself noticed:

the greater the number of sources of data, the better their mantissae fit the law.

Hence if the data X come from various origins and their maxima A come from various origin as well, then both X and A must follow Benford law.

Questions

(a) does there exists a law on $[1, 10[$ followed by both $\mathcal{M}(X)$ and $\mathcal{M}(A)$?

(b) if $\mathcal{M}(A)$ does not verify the same law as $\mathcal{M}(X)$, is it possible to iterate the procedure somehow? Which law do we obtain as a limit?

How to justify Benford's law in terms of the mantissa ?

Many people have wondered why some factors explaining empirical data seem to act multiplicatively.

An interpretation:

we see an everyday-life number X as coming from an interval $[0, A]$, where the maximum A is itself an everyday-life number; this amounts to consider a product, since a continuous random variable on some interval $[0, A]$ can be seen as the product of A by a random variable on $[0, 1]$.

So

Theorem

Let $X = AY$, where Y is a continuous random variable with distribution ν and A is a positive random variable independent of Y. If $\mathcal{M}(A)$ and $\mathcal{M}(X)$ follow the same probability distribution, then this distribution is Benford's law.

This result can be related to the scale-invariance property of Benford's law.

Leading idea:

if there exists a universal law describing the distribution of mantissae of real numbers, it does not depend on the system of measurement. So we expect this law to be scale invariant.

The Theorem naturally leads to consider a Markov chain $(M_n)_{n\geq 1}$, such that M_n follows the same law the mantissa of a product of n independent random variables with law ν.

What is a Markov chain?

A Markov chain is a stochastic model describing a sequence of possible events in which the probability of each event depends on the states attained previously *only through the current state*.

i.e.

If the chain is currently in state s_i, then it moves to state s_j at the next step with a probability which *does not depend upon which states the chain was in before the current one.*

How to justify Benford's law in terms of the mantissa ?

Known fact

Under some conditions, the mantissa of such a product converges to Benford's law.

Indeed we prove

Proposition

The unique invariant measure of M_n is Benford's law.

Meaning:

If we start with M_0 having Benford distribution, then every M_n is Benford.

We also prove that this invariant measure is unique and the convergence is exponential. Precisely

How to justify Benford's law in terms of the mantissa ?

Theorem

$(M_n)_{n\geq 0}$is a Markov chain on [1,10[. Moreover, M_n, conditioned on M_{n-1} has the same law as the mantissa of the product of $M_{n-1}Y$, where Y is an independent random variable with law ν.

Proposition

For every measurable set $B \subseteq [1, 10]$

$$|P(M_n \in B) - \beta(B)| \leq \nu\Big(\Big[\frac{1}{10}, 1\Big]\Big)^n$$

Hence, if $\nu\left(\left[\frac{1}{10}, 1\right]\right) < 1$ the convergence is exponentially fast.

The interest relies in the fact that the exponential speed is expressed in terms of the law ν de Y.

Thank you to the organizers for the invitation

Thank you for your attention