Rank distributions: Frequency vs. magnitude

Carlos Velarde[1◉], **Alberto Robledo**[2◉]*

1 Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, Mexico City, Mexico, **2** Instituto de Física y Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, Mexico City, Mexico

◉ These authors contributed equally to this work.
* robledo@fisica.unam.mx

Abstract

We examine the relationship between two different types of ranked data, frequencies and magnitudes. We consider data that can be sorted out either way, through numbers of occurrences or size of the measures, as it is the case, say, of moon craters, earthquakes, billionaires, etc. We indicate that these two types of distributions are functional inverses of each other, and specify this link, first in terms of the assumed parent probability distribution that generates the data samples, and then in terms of an analog (deterministic) nonlinear iterated map that reproduces them. For the particular case of hyperbolic decay with rank the distributions are identical, that is, the classical Zipf plot, a pure power law. But their difference is largest when one displays logarithmic decay and its counterpart shows the inverse exponential decay, as it is the case of Benford law, or viceversa. For all intermediate decay rates generic differences appear not only between the power-law exponents for the midway rank decline but also for small and large rank. We extend the theoretical framework to include thermodynamic and statistical-mechanical concepts, such as entropies and configuration.

Introduction

Ranking data that originates from apparently disconnected subjects in many fields —astrophysical, geophysical, ecological, biological, technological, financial, urban, social, etc.— has revealed universal patterns [1, 2] and opened intriguing questions about their origin. The empirical law of Zipf [3, 4] for the numbers of occurrence (frequencies if normalized) of words in texts has played a central role in the development of this widespread research topic of multidisciplinary complex systems. Zipf's law has been found to be (approximately) followed by many sets of ranked data outside linguistics, that record the number of occurrences [5] of other types of items. But also, and this is an important distinction we address here, for the magnitudes or sizes of many measurable objects or entities, such as firmament voids, lengths of rivers, city populations, etc. [6].

Here we analyze the conceptual, and also quantitative, difference between frequency and size ranked data. To this purpose we make use of a straightforward stochastic procedure [7–9] to reproduce ranked data from an assumed parent distribution that governs sets of values of random variables that constitute samples. Examination of the expressions for the two types of

rank functions indicate that they are functional inverses of each other. See also [6, 10, 11]. In particular, we focus in the case where the parent distribution $P(N)$, where N is a magnitude random variable, has the power-law form $P(N) \sim N^{-\alpha}$, $1 \leq \alpha < \infty$. We find that in the limit $\alpha = 1$ the size-rank distribution $N(k)$, where k is the rank, decays exponentially as k grows, while the frequency-rank distribution $F(k')$ decays logarithmically as k', the corresponding rank variable, increases. On the contrary, in the limit $\alpha \to \infty$ $N(k)$ decays logarithmically while $F(k')$ does so exponentially. The intermediate case $\alpha = 2$ is the special exponent value when both $N(k)$ and $F(k')$ decay as a power law with exponent -1, the classical Zipf's power law value. To complement our description we replicate the procedure by considering instead a starting parent distribution $Q(F) \sim F^{-\beta}$, $1 \leq \beta < \infty$ where F is a frequency random variable, and obtain an equivalent account with $1 - \beta = 1/(1 - \alpha)$.

We have recently [8, 9, 12] shown that the above-referred stochastic approach to size-rank distributions can be exactly represented by deterministic nonlinear one-dimensional iterated maps close to tangency [13]. Here we extend this strict analogy to determine frequency-size distributions within this nonlinear dynamical language. These distributions are given by areas below map trajectories. To explore the duality between size-rank $N(k)$ and frequency-rank $F(k')$ distributions, we look at specific sets of real data that can be sorted out in both ways, magnitudes or numbers of occurrences, such as the cases of earthquakes [14] and forest fires [15] (see Fig 1), and we find agreement with the theoretical approach. We also comment on how Benford's law [16, 17] for the frequency of digits corresponds in our scheme to the case $\alpha = 1$.

Finally, we extend our statistical-mechanical interpretation with generalized entropies of rank distributions [12, 18] to include the role of $F(k')$.

Rank distributions from a size parent distribution

The basic ingredient in the stochastic method [7–9] for rank distributions is the probability distribution $P(N)$ of the magnitude or size data N under consideration. The scheme is phenomenological since the form of $P(N)$ is assumed, and so, the first common choices are: gaussian, exponential, or power law expressions. For the latter case we write

$$P(N) \sim N^{-\alpha}, \quad 1 \leq \alpha < \infty. \tag{1}$$

Sets of data N can be generated from Eq (1) and subsequently examined if they match, statistically, real ranked data sets. Each data set formed by a total of $\mathcal{N}$ entries, expressed with given suitable precision, can be ranked according to their sizes N or the numbers of times F with which their items appear. We shall consider that N takes positive values within an interval $N_{\min} \leq N \leq N_{\max}$, where we allow as limiting values $N_{\min} = 0$ and/or $N_{\max} \to \infty$. To obtain the number of occurrences F for real numbers N recorded with a given precision it may be necessary to introduce a partition and count incidences within intervals.

The entries in the sample set $\mathcal{N}$ can be sorted out starting with the largest, $N_{\max}$, and continuing with decreasing magnitudes down to $N_{\min}$. And then labeled with the rank variable variable k, with $k = 0$ for $N_{\max}$ and $k = k_{\max}$ for $N_{\min}$. We call the function $N(k)$ the size-rank distribution. The rank k can be an integer $k = 0, 1, 2, 3, \ldots, k_{\max}$ (often, elsewhere, the 1st value is $k = 1$) and it can be generalized to be a real number. The set $\mathcal{N}$ can also be ordered in terms of the frequency with which they appear, that is, the number of occurrences F having size equal or greater than N, or equivalently the rate f, $0 \leq f \leq 1$, of occurrences having size equal or greater than N. For this second sorting the occurrences are labeled with a rank variable k', with $k' = 0$ for the most frequent and $k' = k'_{\max}$ for the least frequent. We call $F(k')$ the frequency-rank distribution. Similarly, the rank k' can be an integer $k' = 0, 1, 2, 3, \ldots, k'_{\max}$ (often the 1st value is $k' = 1$) but it can be generalized to be a real number. The normalized
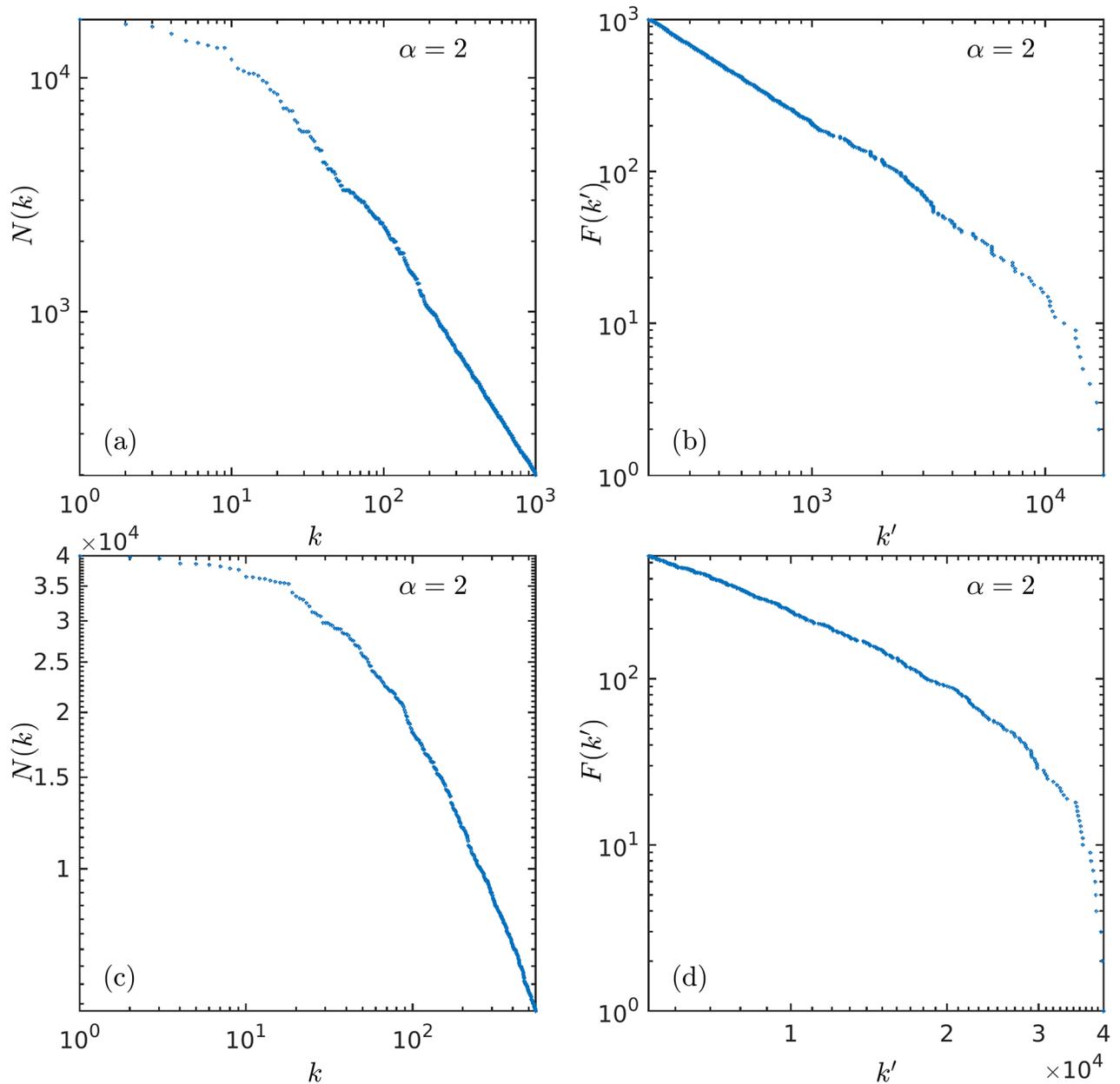
Fig 1. Two examples of ranked data. (a) Data for the energy released by earthquakes in California [14]. (b) Same earthquake data ranked according to number of occurrences of earthquakes of similar magnitude showing behavior compatible with the Guttenberg-Richter law. (c) Data for the areas burnt in forest fires in the U.S.A. [15]. (d) Same forest fires data ranked according to the number of occurrences of similar burnt areas. See text for description.

https://doi.org/10.1371/journal.pone.0186015.g001

frequency-rank distribution is $f(k') = F(k')/\mathcal{N}$. The main task is to determine $N(k)$ and $F(k')$ from $P(N)$.

We now introduce the complementary cumulative distribution of $P(N)$,

$$\Pi(N, N_{\max}) = \int\limits_{N}^{N_{\max}} P(N') dN', \tag{2}$$

where the normalization of $P(N)$ implies $\Pi(N_{\min}, N_{\max}) = 1$. The parent distribution $P(N)$ can be recuperated from $\Pi(N, N_{\max})$ via

$$P(N) = -\frac{\partial}{\partial N}\Pi(N, N_{\max}).$$ (3)

In the theoretical approach the evaluation of $\Pi(N, N_{\max})$ is the means by which the values N generated by $P(N)$ are sorted out and leads to the rank distributions.

The cumulative distribution $\Pi(N, N_{\max})$ increases monotonically as N decreases, taking values from $\Pi(N_{\max}, N_{\max}) = 0$ to $\Pi(N_{\min}, N_{\max}) = 1$. This distribution $\Pi(N(k), N_{\max})$, where we have now indicated the rank k occupied by the variable magnitude N, is identified with $k/\mathcal{N}$, that is

$$\frac{k}{\mathcal{N}} \equiv \Pi(N(k), N_{\max}).$$ (4)

The size-rank distribution $N(k)$ is obtained by solving

$$\frac{k}{\mathcal{N}} = \int\limits_{N(k)}^{N_{\max}} P(N')dN',$$ (5)

for $N(k)$. Normalization of $P(N)$ indicates that $k_{\max} = \mathcal{N}$. If k is to be an integer the possible lower limits in the integral in Eq (5), $N(1), N(2), \ldots, N(k_{\max})$ are such that the integral takes values $1/\mathcal{N}, 2/\mathcal{N}, \ldots, k_{\max}/\mathcal{N}$.

On the other hand, the fraction $k/\mathcal{N}$ can also be seen as the rate or scaled frequency with which the sizes equal or greater than N occur, small for small $k \simeq 0$ and large for $k \simeq k_{\max}$. Therefore we identify the normalized frequency-rank distribution $f(k')$ as

$$f(k') \equiv \Pi(N, N_{\max}),$$ (6)

where $k' \equiv N$. If k' is to be an integer the values of N to be used in $P(N)$ are integers. In practice, the non-normalized frequency-size distribution $F(k') \equiv \mathcal{N}f(k')$ is often used as it is constructed directly from the numbers of occurrences in data samples. From the above definitions $k' \equiv N$, and $F(k') \equiv \mathcal{N}f(k')$, together with Eqs (4) and (6), it is clear that the rank distributions $N(k)$ and $F(k')$ are functional inverses of each other. That is, $k = F(N)$ or $N = F^{-1}(k)$. The inverse of a cumulative distribution is referred to as the quantile function [6, 10]. We refer to $N(k)$ as the size-rank distribution even though technically it is not a probability distribution, as $P(N)$ and $f(k')$ are.

Rank distributions from a power-law parent distribution

We look now at the specific expressions that come out of the general equations in the previous Section when $P(N)$ is given by Eq (1). We have

$$\begin{aligned}
\Pi(N(k), N_{\max}) &= \int\limits_{N(k)}^{N_{\max}} N^{-\alpha}dN \\
&= \frac{1}{1-\alpha}[N_{\max}^{1-\alpha} - N(k)^{1-\alpha}],
\end{aligned}$$ (7)

or, in terms of the q-deformed logarithmic function $\ln_q(x) \equiv (1-q)^{-1}[x^{1-q} - 1]$ with q a real

number,

$$\ln_\alpha N(k) = \ln_\alpha N_{\max} - \mathcal{N}^{-1}k. \tag{8}$$

The size-rank distribution $N(k)$ is explicitly obtained from the above with use of the inverse of $\ln_q(x)$, the q-deformed exponential function $\exp_q(x) \equiv [1 + (1-q)x]^{1/(1-q)}$, this is

$$N(k) = N_{\max}\exp_\alpha[-N_{\max}^{\alpha-1}\mathcal{N}^{-1}k]. \tag{9}$$

While the frequency-rank distribution $f(k')$ is given by

$$\begin{aligned} f(k') &= \ln_\alpha N_{\max} - \ln_\alpha k' \\ &= 1 + \ln_\alpha N_{\min} - \ln_\alpha k'. \end{aligned} \tag{10}$$

In Fig 2 we show the agreement of Eqs (9) and (10) with the data on earthquakes and forest fires already shown in Fig 1. Our method for fitting the data to Eqs (9) and (10) is heuristic. We first select a data point to define $N_{\max}$. We then approximate with a straight line segment a section of the data that appears lined when displayed in logarithmic scales (involving a choice of its two extremes) via minimum squares. This gives us, with the use of Eq (8), a set of two equations from which we determine numerically preliminary values for α and $\mathcal{N}$ (notice that Eq (7) has no normalization constant). We iterate this procedure to improve fitting (mostly only $\mathcal{N}$ changes its value appreciably). Once the parameters in Eq (9) are determined $F(k')$ follows from Eq (10).

When $\alpha = 1$ Eq (9) acquires the ordinary exponential form

$$N(k) = N_{\max}\exp(-\mathcal{N}^{-1}k), \tag{11}$$

while Eq (10) becomes an ordinary logarithmic function,

$$\begin{aligned} f(k') &= \ln(N_{\max}/k') \\ &= 1 - \ln(k'/N_{\min}). \end{aligned} \tag{12}$$

We take the limit $\alpha \to \infty$ to signify that $P(N) = N_0\exp(-N_0 N)$, and we choose $N_0 = 1$. We find

$$N(k) = -\ln[\exp(-N_{\max}) + \mathcal{N}^{-1}k], \tag{13}$$

and

$$f(k') = \exp(-k') - \exp(-N_{\max}) \tag{14}$$

$$= \exp(-k') - \exp(-N_{\min}) + 1. \tag{15}$$

In the limit $N_{\max} \to \infty$ Eq (9) becomes the power law $N(k) \sim k^{1/(1-\alpha)}$ that when $\alpha = 2$ gives the simple hyperbolic form $N(k) \sim k^{-1}$. Whereas Eq (10) in the same limit becomes the power law $f(k') \sim k'^{(1-\alpha)}$ that when $\alpha = 2$ gives, coincidentally, the same hyperbolic form $f(k') \sim k'^{-1}$. For many sets of frequency-rank real data $\alpha \simeq 2$ and the standard Zipf law is $\alpha = 2$, whereas the same feature for real size-rank data has led to refer (concurrently) to the observation of Zipf's law in relation to $N(k)$. In contrast, when $\alpha \to \infty$, in the limit $N_{\max} \to \infty$ the rank distributions become $N(k) = \ln(\mathcal{N}/k)$ and $f(k') = \exp(-k')$, $N(k)$ decays very fast as k increases since the argument in the logarithmic function lies in the interval $0 < k/\mathcal{N} < 1$, while $f(k')$ decays exponentially as k' increases. This can be compared with the case $\alpha = 1$, but $N_{\max}$ finite, when $N(k)$ decays exponentially as k increases while $f(k')$ decays very fast as k'
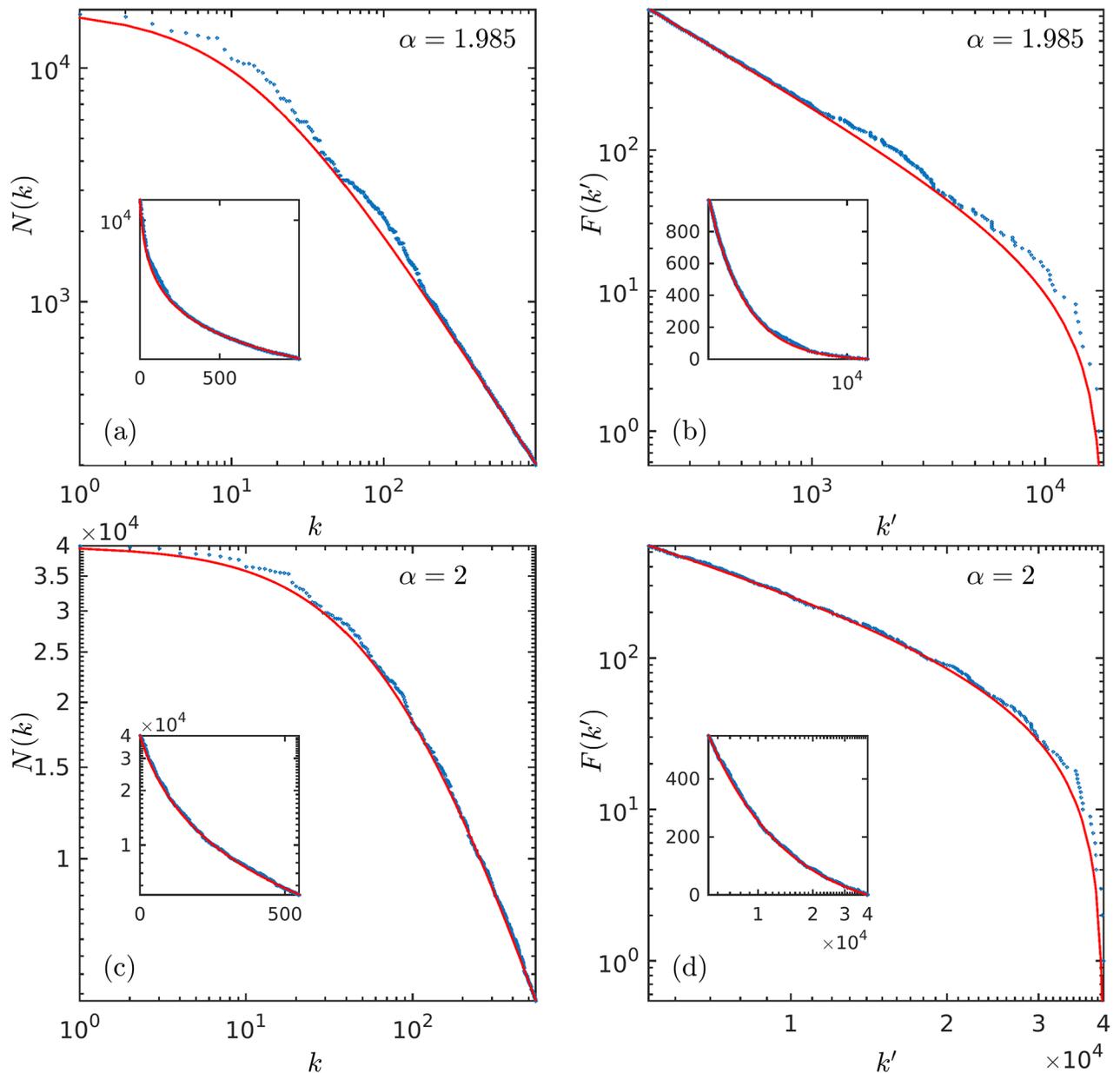
Fig 2. Theoretical fitting of ranked data. Same two examples in Fig 1 of ranked data on earthquakes and forest fires fitted with the expressions in Eqs (9) and (10). (a) Size-rank distribution $N(k)$ for earthquakes. (b) Frequency-rank distribution for earthquakes (with $F(k') = \mathcal{N}f(k')$). (c) Size-rank distribution $N(k)$ for forest fires. (d) Frequency-rank distribution for forest fires (with $F(k') = \mathcal{N}f(k')$). As can be seen, the values of α needed for fitting are close to $\alpha = 2$ that corresponds to the classical Zipf law exponent. See text for description.

increases since again the argument in the logarithmic function lies in the interval $0 < k'/N_{\max} < 1$.

A note on normalization. The choice of $P(N)$ given by Eq (1) is not compatible with finite data sets ($\mathcal{N} < \infty$), these should be represented by a different expression for $P(N)$, at least one that differs from Eq (1) for some values of N, specially small N. Normalization of Eq (1) obeys $k_{\max} = \mathcal{N}$, with both $k_{\max} \to \infty$ and $\mathcal{N} \to \infty$, while $N_{\min} \to 0$.

Rank distributions from a frequency parent distribution

To show a duality feature of the approach to rank distributions we now consider the derivation of these distributions from a different parent distribution. This distribution, $Q(F)$, generates values of the numbers of occurrences F to form data sets. As before we introduce a complementary cumulative distribution

$$X(F, F_{max}) = \int_F^{F_{max}} Q(F')dF', \tag{16}$$

where the normalization of $Q(F)$ implies $X(F_{min}, F_{max}) = 1$. We denote by $\mathcal{F}$ the total number of elements in the occurrences sample set.

Proceeding as before we indicate the rank k' occupied by the number of occurrences F in the distribution $X(F(k'), F_{max})$ and identify this as $k'/\mathcal{F}$. That is

$$\frac{k'}{\mathcal{F}} \equiv X(F(k'), F_{max}). \tag{17}$$

When we assume the power law expression

$$Q(F) \sim F^{-\beta}, \;\; 1 \leq \beta < \infty, \tag{18}$$

we obtain

$$X(F(k'), F_{max}) = \frac{1}{1 - \beta}[F_{max}^{1-\beta} - F(k')^{1-\beta}], \tag{19}$$

or, in terms of the q-deformed logarithmic function,

$$\ln{}_\beta F(k') = \ln{}_\beta F_{max} - \mathcal{F}^{-1}k'. \tag{20}$$

The frequency-rank distribution $F(k')$ is explicitly obtained from the above with with use of the q-deformed exponential function, this is

$$F(k') = F_{max} \exp{}_\beta[-F_{max}^{\beta-1}\mathcal{F}^{-1}k']. \tag{21}$$

While the size-rank distribution $N(k)$, following arguments parallel to those given before for $F(k')$, is given by

$$N(k) \equiv \mathcal{F}X(k, F_{max}), \tag{22}$$

where $k \equiv F$. Explicitly,

$$\begin{aligned} N(k) &= \ln{}_\beta F_{max} - \ln{}_\beta k \\ &= 1 + \ln{}_\beta F_{min} - \ln{}_\beta k. \end{aligned} \tag{23}$$

Again, it is clear that the rank distributions $F(k')$ and $N(k)$ are functional inverses of each other. That is, $k' = N(F)$ or $F = N^{-1}(k')$.

The exponent α in the previous two sections and the exponent β in this section are related via

$$1 - \alpha = \frac{1}{1 - \beta}, \tag{24}$$

and coincide in value when $\alpha = \beta = 2$, and both distributions acquire the simple hyperbolic

functions $F(k') \sim k'^{-1}$ and $N(k) \sim k^{-1}$ when in addition $F_{\max} \to \infty$ and $N_{\max} \to \infty$, that is, the classical Zipf case.

Rank distributions from a nonlinear map at tangency

We have shown recently [8, 9, 12] that there is an exact analogy between the expressions for the rank distributions as presented above for $N(k)$ and those for the trajectories associated with the tangent bifurcation in one-dimensional nonlinear iterated maps. A map $g(x)$ at the tangent bifurcation is written locally as $x' = g(x) = x - u|x|^z + \cdots, x \leq 0$ [13], $z > 1$, and trajectories initiated at $x_0 \lesssim 0$ are obtained via repeated iterations of $g(x)$, i.e.

$$x_{\tau+1} = x_\tau - u|x_\tau|^z, \tau = 0, 1, \ldots \tag{25}$$

These trajectories move monotonically towards the point of tangency at $x = 0$. If we make the replacement, valid for large time τ, of the difference $x_{\tau+1} - x_\tau$ by $dx_\tau/d\tau$ in Eq (25) (written as $-u|x_\tau|^z = x_{\tau+1} - x_\tau$) we obtain the differential form $ud\tau = -|x_\tau|^{-z} dx_\tau$, and integration of both sides of it yields

$$
\begin{aligned}
ut &= \int_{x_0}^{x_t} \frac{dx_\tau}{-|x_\tau|^z} \\
&= \frac{1}{1-z}[-|x_t|^{1-z} + |x_0|^{1-z}],
\end{aligned}
\tag{26}
$$

or

$$\ln_z|x_t| = \ln_z|x_0| - ut. \tag{27}$$

The iteration number or time t dependence of all trajectories is obtained by solving the above for x_t, i.e.

$$x_t = x_0 \exp_z[-|x_0|^{1-z}ut]. \tag{28}$$

The equivalence of the trajectory positions x_t with the size-rank distribution $N(k)$ is made clear by comparison of Eqs (27) and (28) with Eqs (8) and (9), respectively, together with the identifications $t = k$, $u = \mathcal{N}^{-1}$, $x_t = -N(k)$, $x_0 = -N_{\max}$ and $z = \alpha$. Also, comparison of the right-hand side of Eq (26) with that of Eq (7), taking into account Eq (6), indicates that the analog of the frequency-rank distribution $f(k')$ is the quantity

$$
\begin{aligned}
A_t &= \int_{x_0}^{x_t} \frac{dx_\tau}{-|x_\tau|^z} \\
&= \ln_z|x_0| - \ln_z|x_t|,
\end{aligned}
\tag{29}
$$

where $-x_t$ plays the role of k'. In [8] it is pointed out that the trajectories given by Eq (28) have precisely the analytical form for all trajectories with generic x_0 that are generated by the functional composition renormalization group fixed-point map [13, 19] at the tangent bifurcation. And therefore the areas A_t in Eq (29) have also the same property. That is, all trajectories of the fixed-point map for all t initiated at the generic position x_0 obey Eq (28). Also Eq (29) enjoys the degree of universality given by the fixed-point map.

In Fig 3 we illustrate the iterated map properties for the case $z = \alpha = 2$ that translate into the equivalent description of the rank distributions $N(k)$ and $F(k')$.

When $z = 1$ we have

$$x_t = x_0 \exp[-ut]. \tag{30}$$
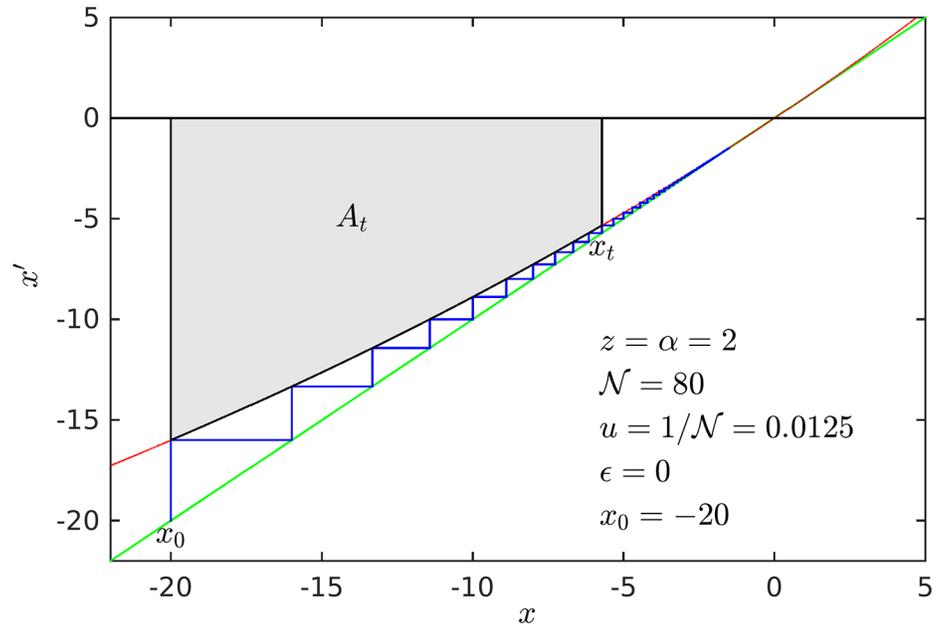
Fig 3. Iterated map at tangency. The map parameters are $z = \alpha = 2$, the curvature is $u = 0.0125$, and the trajectory x_t, $t = 0, 1, 2, \ldots t$, is initiated at x_0 as given by Eq (28). Also shown is the area A_t (shaded) as given by Eq (29). The map properties translate into the equivalent description of the rank distributions $N(k)$ and $F(k')$ via the identifications $t = k$, $u = \mathcal{N}^{-1}$, $x_t = -N(k) = k'$, $x_0 = -N_{max}$ and $A_t = F(k'/\mathcal{N})$. See text for description.

and

$$A_t = \ln |x_t| - \ln |x_0|. \tag{31}$$

The trajectories in Eq (30) are obtained when a linear map intersects the identity line, i.e.

$$x' = f(x) = (1 - a)x, \tag{32}$$

and this occurs locally when the tangent map is shifted into a double-secant map.

In the limit $z \to \infty$ the counterpart of Eq (28) is

$$x_t = \ln [\exp(x_0) + ut], \tag{33}$$

as this expression transforms into Eq (13) for $N(k)$ under the same equivalences $t = k$, $u = \mathcal{N}^{-1}$, $x_t = -N(k)$, $x_0 = -N_{max}$, while that corresponding to Eq (29) is

$$A_t = \exp(x_0) - \exp(x_t). \tag{34}$$

Rank distributions associated with Benford's first digit law

Benford's first digit law [16, 17],

$$\pi(n) = \log \left(\frac{n + 1}{n} \right), \tag{35}$$

where n is the first digit of a decimal base number N and log denotes the decimal base logarithmic function, can be readily expressed in terms of the complementary cumulative distribution
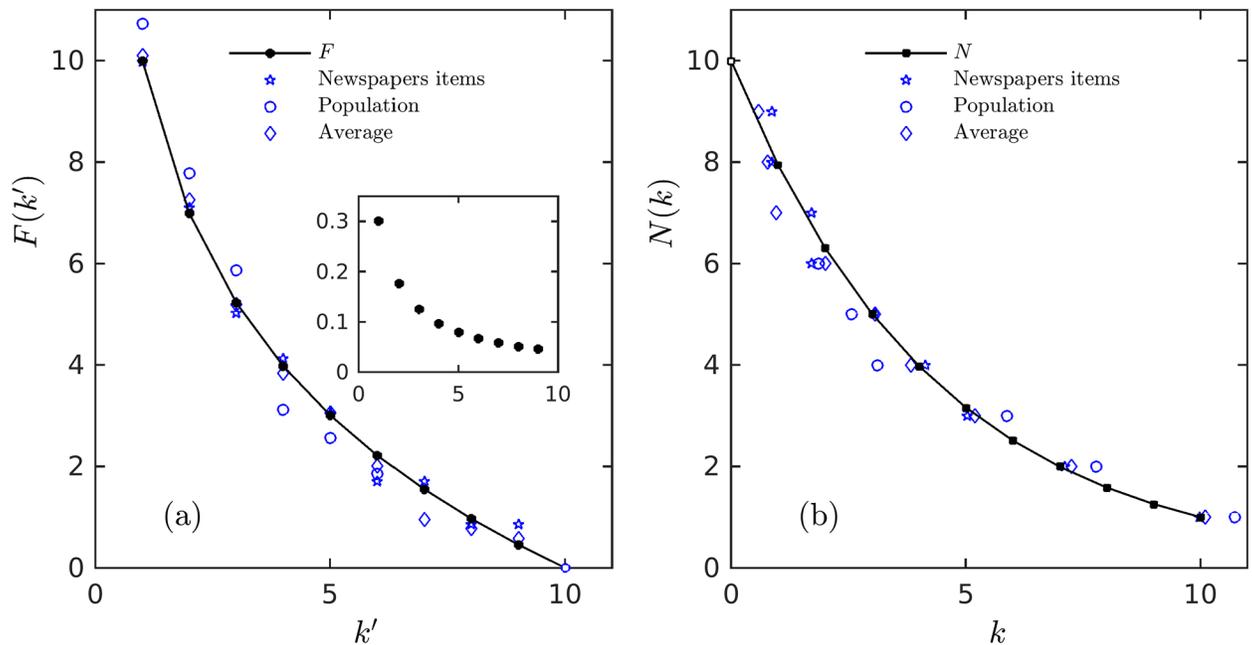
Fig 4. Rank distributions for Benford law. Rank distributions for Benford law together with numerical data that follows this law (shown in the inset). (a) Frequency-rank distribution $F(k')$. (b) Size-rank distribution $N(k)$. They are obtained from the general formalism with $\alpha = 1$. Data taken from Table I in the original Benford's article Ref. [17]. See text for description.

Eq (2) when $\alpha = 1$ and the parent distribution is $P(N) = 1/N$. This is

$$\pi(n) = \Pi(N, N_{\max}) - \Pi(N + 1, N_{\max}) = \log(n + 1) - \log(n),\qquad(36)$$

where $N + 1 = (n + 1).000\cdots$ and $N = n.000\cdots$.

Thus, by considering the cumulative version of Benford's law,

$$\Pi(N, N_{\max}) = \log(N_{\max}) - \log(N),\qquad(37)$$

with $N_{\max} = 10$ and $N = n.000\cdots$, $n = 1, 2, \ldots, 9$, we have

$$F(k') = \mathcal{N} \log(N_{\max}/k'), k' = 1, 2, \ldots, N_{\max},\qquad(38)$$

and

$$N(k) = N_{\max} 10^{-k/\mathcal{N}}, k = 0, 1, \ldots, N_{\max}.\qquad(39)$$

In Fig 4 we show these distributions together with numerical data that follows Benford's law as shown in the figure's inset.

Benford's law has been generalised to the case $\alpha > 1$ [7], so that its associated (complementary) cumulative distribution Eq (7) provides the connection with the rank distributions studied here. In particular the case $\alpha = 2$ corresponds to the classical Zipf's law described by $F(k')$ with $k' = 0, 1, 2, 3, \ldots$, shifted and limited to the values of the first digits $1, 2, 3, \ldots, 9$, when using decimal base logarithms.

Rank distributions as expressions of a thermodynamic structure

As pointed out $N(k)$ is a the functional inverse of $F(k')$, that is, the inverse of a (non-normalized complementary) cumulative distribution in reverse order, a quantile function [6, 10]. Also $N(k)$ has been interpreted [8, 9] as the total number that the size variable *unit* occurs at fixed rank k. That is, in thermal system language, $N(k)$ is equivalent to the degeneracy of a micro state of 'energy' k, or a micro-canonical partition function with fixed k, where the associated uniform probability is $p_i^{(k)} = p^{(k)} \equiv 1/N(k)$ for all $i = 1, \ldots, N(k)$. Thus, we can call $S(k) \equiv \ln N(k)$ an entropy for $\alpha = 1$ and define $S(k) \equiv \ln_\alpha N(k)$ as a generalized entropy for $\alpha > 1$. Likewise $S_{\max} \equiv \ln N_{\max}$ for $\alpha = 1$ and $S_{\max} \equiv \ln_\alpha N_{\max}$ for $\alpha > 1$, $N(0) \equiv N_{\max}$. Eq (8) is written now as

$$S_{\max} = S(k) + \mathcal{N}^{-1} k, \tag{40}$$

where, if $S(k)$ is thought of as the entropy for the system with fixed k, then $S_{\max}(\mathcal{N}^{-1})$ would be a generalized Massieu potential when the variable k is replaced by the (conjugate) variable $\mathcal{N}^{-1}$ via a Legendre transformation.

Just like thermodynamic quantities are dominant values of statistical-mechanical fluctuating quantities in a macroscopic system, we think of Eq (11), valid for $\alpha = 1$,

$$N_{\max} = N(k) \exp\left(\mathcal{N}^{-1} k\right), \tag{41}$$

to be the result of the application of the saddle-point approximation for large $\mathcal{N}$ on

$$N_{\max} = \int N(k') \exp\left(\mathcal{N}^{-1} k'\right) dk'. \tag{42}$$

The consideration of the emergence of dominant rank fluctuations for the general case $\alpha > 1$ in the 'thermodynamic' limit $k_{\max} = \mathcal{N} \to \infty$ is less straightforward and here we do not discuss it further.

We recall [20, 21] that the formalism of thermodynamics can be expressed in two equivalent ways. One of them is to consider as starting point the entropy as the fundamental monotonic function of the energy (and other basic variables) that characterise the system, while the other alternative is to begin with the internal energy as the fundamental quantity, a monotonic function of the entropy (and the same other variables). The expressions obtained from the parent distribution $P(N)$, Eqs (8) and (9), correspond to the former choice, while those obtained from the parent distribution $Q(F)$, Eqs (20) and (21), relate to the second one.

In this statistical-mechanical interpretation the rank k plays the role of energy and the entropy is $S(k) = \ln N(k)$, when $\alpha = 1$. In the alternative description F plays the role of energy while the entropy is $\ln(k')$, again when $\alpha = 1$. Thus here the quantities representing entropy and energy are, as customary, functional inverses of each other in accordance with the usual two equivalent thermodynamic frames [20, 21].

As a consequence of the precise analogy between the rank distributions obtained from a parent distribution and the nonlinear iterated fixed-point map at tangency, we note that the thermodynamic structure observed above for the rank distributions quantities translates thoroughly into an equivalent structure for the nonlinear dynamical problem. It is only necessary to recall the identifications $z = \alpha$, $t = k$, $u = \mathcal{N}^{-1}$, $x_t = -N(k)$, $x_0 = -N_{\max}$ and $A_t = -f(k')$ with $x_t = -k'$.

Summary and discussion

We have analyzed the relationship that exists between two types of ranked data, numbers of occurrences and sizes or magnitudes of items. The technical relationship is well understood

for statistics specialists, frequency-rank data is represented by a (complementary) cumulative probability distribution while size-rank data is described by its functional inverse, a quantile function [6, 10]. It is of wider interest, for those studying the many topics of the complex systems science, where universal patterns are observed in ranked data samples from very different sources, such as the empirical laws of Zipf and Benford, to understand the physical origin of the documented behavior We have obtained expressions for size-rank $N(k)$ and frequency-rank $F(k')$ distributions from a stochastic method and or from an equivalent nonlinear deterministic approach and corroborated that the two functions are inverses of each other. Their differences are most apparent when the exponent α of the power-law parent distribution differs from $\alpha = 2$, but they coincide and behave as hyperbolic functions (with deviations for small and large rank) when $\alpha = 2$. In this latter case we have a nonlinear map at tangency with nonzero curvature, the most common case of analytic map at tangency. This being the case of the classic Zipf law. On the other hand we illustrated the case when $\alpha = 1$ with the first digit Benford law.

We complemented our description by also considering the option for the parent distribution for the source of data to be that for the number of occurrences F instead of that for the size N. When these two distributions are assumed to have the power-law forms $P(N) \sim N^{-\alpha}$ and $Q(F) \sim F^{-\beta}$ we obtain parallel (and equivalent) descriptions for the size and frequency rank distributions with the roles of cumulative distribution and quantile function interchanged and with the exponents relationship $1 - \alpha = (1 - \beta)^{-1}$. Further, we advanced a thermodynamic and statistical-mechanical interpretation to be associated with the properties obtained for the rank distributions and indicated that the rank k plays the role of energy and $N(k)$ takes the place in a prototypical thermal system of the number of configurations at fixed energy k with entropy $S(k) = \ln N(k)$ when $\alpha = 1$. The interpretation of the alternative description corresponds to F playing the role of energy while $\ln(k')$ that of entropy when $\alpha = 1$. Thus entropy and energy as functional inverses of each other provide two equivalent thermodynamic formalisms [20, 21].

The case $\alpha > 1$ suggests the use of the generalized entropy expression $S(k) = \ln_\alpha(k)$ in the thermodynamic description, but this poses a question for its corresponding statistical-mechanical formalism in that the validity of the usual saddle-point approximation requires reconsideration. The outcome may be one in which fluctuations are not suppressed in the thermodynamic limit, here represented by $k_{\max} \to \infty$ and $\mathcal{N} \to \infty$. The reproduction of the rank distributions via a nonlinear map at a tangent bifurcation indicates a reason for the appearance of generalized entropy expression through the drastic contraction of configuration space from a real number set of possible iterated map trajectories positions to only a finite number in the limit $k_{\max} \to \infty$ and $\mathcal{N} \to \infty$ that corresponds to $t \to \infty$ [12, 18].

Author Contributions

Conceptualization: Carlos Velarde, Alberto Robledo.

Formal analysis: Carlos Velarde, Alberto Robledo.

Funding acquisition: Alberto Robledo.

Investigation: Carlos Velarde, Alberto Robledo.

Methodology: Carlos Velarde, Alberto Robledo.

Writing – original draft: Carlos Velarde, Alberto Robledo.

Writing – review & editing: Carlos Velarde, Alberto Robledo.

References

1. To Honor G.K. Zipf Glottometrics 3,4,5 Ludenscheid: RAM-Verl., 2002 ISSN 1617-8351. Available from: http://www.ram-verlag.eu/journals-e-journals/glottometrics/

2. Newman M.E.J. Power laws, Pareto distributions and Zipf's law. Contemporary Physics 2005; 46 (5);323–351. https://doi.org/10.1080/00107510500052444

3. Zipf's law. Wikipedia. Available from: https://en.wikipedia.org/wiki/Zipf

4. Zipf GK. Human Behavior and the Principle of Least Effort. Cambridge: Addison-Wesley, 1949.

5. Frequency-rank distributions. Wikipedia. Available from: https://en.wikipedia.org/wiki/Cumulative_frequency_analysis

6. Rank-size distributions. Wikipedia. Available from: https://en.wikipedia.org/wiki/Rank-size_distribution

7. Pietronero L. Tosatti E. Tosatti V. Vespignani A. Explaining the uneven distribution of numbers in nature: the laws of Benford and Zipf. Physica A. 2001; 293(1-2):297–304. https://doi.org/10.1016/S0378-4371(00)00633-6

8. Altamirano C. Robledo A. Possible thermodynamic structure underlying the laws of Zipf and Benford. Eur Phys J B. 2011; 81(3):345–351. https://doi.org/10.1140/epjb/e2011-10968-5

9. Robledo A. Laws of Zipf and Benford, intermittency, and critical fluctuations. Chinese Sci Bull. 2011; 56 (34):3645–3648. https://doi.org/10.1007/s11434-011-4827-y

10. Quantile function. Wikipedia. Available from: https://en.wikipedia.org/wiki/Quantile_function

11. Egghe L. Waltman L. Relations between the shape of a size-frequency distribution and the shape of a rank-frequency distribution Information Processing and Management 2011; 47:238–245. https://doi.org/10.1016/j.ipm.2010.03.009

12. Yalcin G.C. Robledo A. Gell-Mann M. Incidence of q statistics in rank distributions. Proc Natl Acad Sci USA. 2014; 111(39):14082–14087. https://doi.org/10.1073/pnas.1412093111 PMID: 25189773

13. Schuster H.G. Deterministic Chaos. An Introduction. VCH Publishers, Weinheim; 1988.

14. Southern California Earthquake Data Center. Available from https://www.data.scec.org

15. Forest Fires Data. Clauset home page Data and Code. Santa Fe Institute. Available from: http://tuvalu.santafe.edu/∼aaronc/powerlaws/data.htm

16. Benford's law. Wikipedia. Available from: https://en.wikipedia.org/wiki/Benford'slaw

17. Benford F. The law of anomalous numbers. Proc Am Phil Soc, 1938; 78: 551–572.

18. Yalcin G.C. Velarde C. Robledo A. Generalized entropies for severely contracted configuration space. Heliyon, 2015; 1:e00045. Available from: http://www.heliyon.com/article/e00045/ https://doi.org/10.1016/j.heliyon.2015.e00045 PMID: 27441229

19. Hu B. Rudnick J. Exact solutions to the Feigenbaum renormalization- group equations for intermittency. Phys Rev Lett, 1982; 48: 1645–1648. https://doi.org/10.1103/PhysRevLett.48.1645

20. Tisza L. Generalized thermodynamics. Cambridge University Press, Cambridge; 1993.

21. Callen H.B. Thermodynamics and an introduction to thermostatistics. Wiley, Cambridge; 1993.