

Generalized Thermodynamics Underlying the Laws of Zipf and Benford

Carlo Altamirano and Alberto Robledo

Instituto de Física, Universidad Nacional Autónoma de México,
Apartado postal 20-364, México 01000 D.F., México

Abstract. We demonstrate that the laws of Zipf and Benford, that govern scores of data generated by many and diverse kinds of human activity (as well as other data from natural phenomena), are the centerpiece expressions of a generalized thermodynamic structure. This structure is obtained from a deformed type of statistical mechanics that arises when configurational phase space is incompletely visited in an especially severe fashion. Specifically, the restriction is that the accessible fraction of this space has fractal properties. We obtain a generalized version of Benford's law for data expressed in full and not by the first digit. The inverse functional of this expression is identified with the Zipf's law; but it naturally includes the tails observed in real data for small rank. Thermodynamically, our version of Benford's law expresses a Legendre transform between two entropy (or Massieu) potentials, while Zipf's law is merely the expression that relates the corresponding partition functions.

Keywords: Zipf's law, Benford's law, generalized thermodynamics, fractal phase space.

1 Introduction

Over more than half a century, observers of the astonishing ubiquity of the empirical laws of Zipf and Benford have been puzzled by their seeming universal validity and intrigued about the plausible answer to the central question of why they appear in so many contexts. As it is widely known, Zipf's law refers to the (approximate) power law that is displayed by sets of data (populations of cities, words in texts, impact factors of scientific journals, etc.) when these are given a ranking (in relation to size of populations, frequency of words, magnitude of impact factors, etc.) [1]. Benford's law is a well-known simple logarithmic rule for the frequency of first digits found in listings of data (stock market prices, census data, heat capacities of chemicals, etc.) [2].

It has been argued that Benford's law is a special case of Zipf's law [3]. Indeed the relationship between the two has been made explicit some years ago [4] by first obtaining a generalization of Benford's law from the basic assumption that the underlying probability distribution $P(N)$ of the data N under consideration is scale invariant and therefore has the form of the power law $P(N) \sim N^{-\alpha}$,

$\alpha > 0$. A simple integration over $P(N)$, to obtain the relative probability for consecutive integers n and $n + 1$, leads, when $\alpha = 1$, to $P(n) = \log(1 + n^{-1})$ which is Benford's law. The next step in Ref. [4] was to obtain the rank k from $P(N)$, this time as an integration over $P(N)$ from $N(k)$, the number of data that define the rank k , to a finite number N_{\max} that corresponds to rank $k = 1$. In the limit $N_{\max} \rightarrow \infty$ they obtain $N(k) \sim k^{1/(1-\alpha)}$ that is Zipf's law with exponent $1/(1 - \alpha)$ when $\alpha > 1$.

Here we expand on the results of Ref. [4] merely by keeping N_{\max} finite, and as we argue below, this consideration facilitates the articulation of a major inference on the physical nature of the laws of Zipf and Benford. We contend that these laws represent general thermodynamic relations, albeit for a special type of thermodynamic structure obtained from the usual via a deformation parameter. The generalized form of the Benford's law as derived in Ref. [4] in its broad form (not specialized to a first or other digit) is seen to express a Legendre transform between two thermodynamic potentials, the inverse of which becomes a generalized Zipf's law. We also reason that this kind of deformed thermodynamics arises from the introduction of a strong impediment in accessing configurational phase space that results in a fractal or multifractal available subset of this space. We identify the quantities that represent thermodynamic potentials, partition functions and conjugate thermodynamic variables. In the next Section 1 we reproduce the results in Ref. [4] relevant to our purposes and then describe in the following Section 2 our main results. We conclude the article with a short Summary and discussion.

2 Derivation of the Laws of Benford and Zipf

Denote by $P(N)$ the probability distribution associated to the set of data under consideration (e.g. the distribution obtained from a histogram generated by data - a total of \mathcal{N} numbers - giving the magnitudes of the population of a set of countries). Under the assumption of scale invariance the distribution has the form of a power law $P(N) \sim N^{-\alpha}$. The probability of observation of the first digit n of number N is given by [4]

$$P(n) = \int_n^{n+1} N^{-\alpha} dN = \frac{1}{1 - \alpha} [(n + 1)^{1-\alpha} - n^{1-\alpha}], \quad \alpha \neq 1, \tag{1}$$

from which one obtains Benford's law $P(n) = \log(1 + n^{-1})$ when $\alpha = 1$.

The set of \mathcal{N} factual data numbers can be ranked and compared with ranking of another set of also \mathcal{N} numbers extracted from the basic distribution $P(N) \sim N^{-\alpha}$. The rank k is given by [4]

$$k = \mathcal{N} \int_{N(k)}^{N_{\max}} N^{-\alpha} dN = \frac{1}{1 - \alpha} [N_{\max}^{1-\alpha} - N(k)^{1-\alpha}], \quad \alpha \neq 1, \tag{2}$$

where N_{\max} and $N(k)$ correspond, respectively, to rank $k = 1$ and nonspecific rank $k > 1$. Inversion of the above in the limit $N_{\max} \gg 1$ yields Zipf's law $N(k) \sim k^{1/(1-\alpha)}$.

3 Laws of Benford and Zipf and q -Deformed Thermodynamics

Consider the q -deformed logarithmic function

$$\log_q(x) \equiv \frac{1}{1-q} [x^{1-q} - 1], \tag{3}$$

with $q \neq 1$ a real number, and its inverse, the q -deformed exponential function

$$\exp_q(x) \equiv [1 + (1-q)x]^{1/(1-q)}, \tag{4}$$

that reduce, respectively, to the ordinary logarithmic and exponential functions when $q = 1$. In terms of these functions, Eq. (2) and its inverse can be written more economically as

$$\mathcal{N}^{-1}k = \log_\alpha N_{\max} - \log_\alpha N(k), \tag{5}$$

and

$$N(k) = N_{\max} \exp_\alpha [-N_{\max}^{\alpha-1} \mathcal{N}^{-1}k]. \tag{6}$$

We first comment that Eq. (6) is a generalization of Zipf's law that takes into account, as we see below, the behavior for low rank k observed in real data where N_{\max} is finite. Clearly, we recover from Eq. (6) the power law $N(k) \sim k^{1/(1-\alpha)}$ in the limit $N_{\max} \gg 1$ when $\alpha > 1$.

Now, in order to arrive at an interesting physical interpretation of Eq. (5) we look at the quantities contained in it. We notice that both $\log_\alpha N_{\max}$ and $\log_\alpha N(k)$ are given by the integrals

$$\log_\alpha N_{\max} = \int_1^{N_{\max}} N^{-\alpha} dN \text{ and } \log_\alpha N(k) = \int_1^{N(k)} N^{-\alpha} dN, \tag{7}$$

and these in turn can be seen, when $\alpha = 1$, to conform to the evaluation of entropy $\widehat{S}_1 = \log N_{\max}$ or $S_1 = \log N(k)$ where the probability of N equally probable configurations in phase space is $P(N) = N^{-1}$ [5]. If we now allow for $\alpha > 1$ we can retain the same interpretation,

$$\widehat{S}_\alpha = \log_\alpha N_{\max} \text{ and } S_\alpha = \log_\alpha N(k), \tag{8}$$

with $P(N) = N^{-\alpha}$, with N_{\max} and $N(k)$ playing the roles of total configurational numbers or partition functions. Therefore Eq. (5) can be rewritten as

$$S_\alpha = \widehat{S}_\alpha - \mathcal{N}^{-1}k,$$

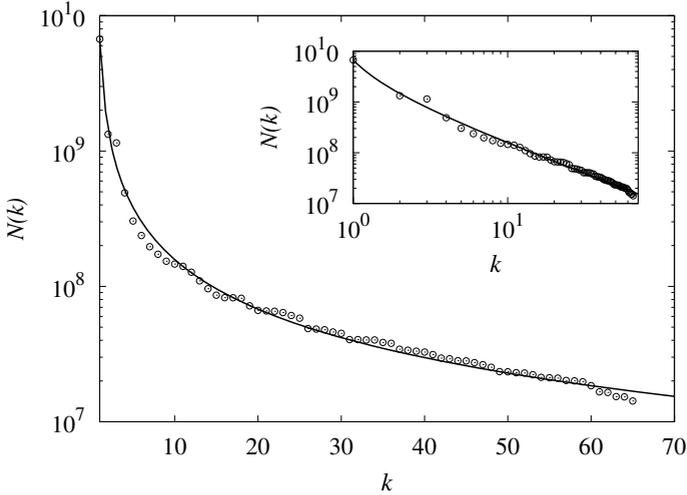


Fig. 1. Rank-order statistics for the world population by country (*empty circles*) taken from [6]. The x axis represents the rank while the y axis stands for the population. An $\exp_\alpha(x)$ function with $\alpha \simeq 1.86$ (*smooth curve*) is fitted to the data.

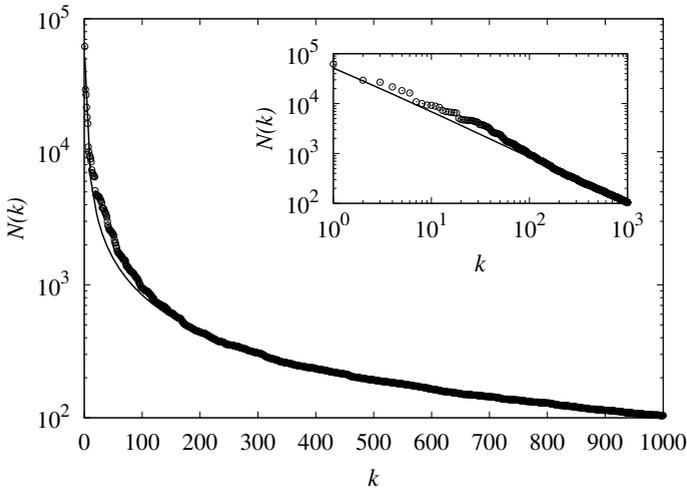


Fig. 2. Rank-order statistics for the appearance of words (*empty circles*) in the British National Corpus [7]. The x axis represents the rank while the y axis stands for the frequency of appearance of each word. An $\exp_\alpha(x)$ function with $\alpha \simeq 2.09$ (*smooth curve*) is fitted to the data.

and read as the expression of a Legendre transform from the Massieu potential $\widehat{S}_\alpha(\mathcal{N}^{-1})$, a function of the inverse of the number \mathcal{N} , to the entropy $S_\alpha(k)$, a function of the rank k . The conjugate variables \mathcal{N}^{-1} and k , could be seen to play the roles of inverse temperature β and energy u in the description of a thermal system.

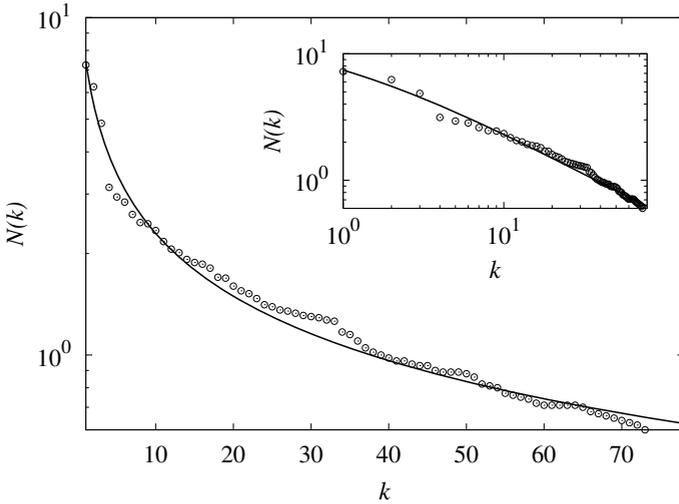


Fig. 3. Rank-order statistics for journals by impact factor (*empty circles*) [8]. The x axis represents the rank while the y axis stands for the impact factor value of each journal. An $\exp_{\alpha}(x)$ function with $\alpha \simeq 2.53$ (*smooth curve*) is fitted to the data

Eq. (6), being the inverse of Eq. (5), states the same relationship but in terms of the partition functions $N(k)$ and $N_{\max}(\mathcal{N}^{-1})$. We illustrate in Figs. 1 to 3 the capability of Eq. (6) to reproduce real data for rankings of city populations, words in texts and impact factors of scientific journals, respectively.

4 Summary and Discussion

We have presented here a novel thermodynamic, statistical-mechanical, reading of the generalized law of Benford and its inverse, an also generalized law of Zipf. This interpretation may explain the ever presence of these phenomenological laws in a wide range of observations including very dissimilar situations. We remark here that the deviation from unity of the exponent α implies a restricted access to the phase space values for N , this restriction involves an accessible subset of this space with a scale invariant property, i.e. a fractal set, as stated by the power law $P(N) \sim N^{-\alpha}$. Our identification of the rank k as analogous to an energy u is supported by the fact that the maximum value of N , N_{\max} , can be replaced by any other reference value $N_{ref} < N_{\max}$ in the derivation of Eq. (2) (and therefore of Eqs. (5) and (6)) that results in a shift in the interval of rank values as they would start now at a negative k . This interval can be shifted again to recover $k = 1$ as the highest rank. Such shift would correspond to a common shift in the interval of energy values u . On the other hand the identification of \mathcal{N}^{-1} with the inverse temperature β indicates that the larger the sample \mathcal{N} the smaller the ‘temperature’. As we observe in Figs. 1 to 3, Eq. (6) succeeds in reproducing the small rank tail present in real data that appears

before the power law behavior sets in, and it is clearly due to the finite size described by $N_{\max} < \infty$. Real data also show deviations from the power law $N(k) \sim k^{1/(1-\alpha)}$ for large values of the rank k [9] [10]. We do not address these deviations here but we discuss them in a future publication [11] where we make use of an unstated duality property of rank functions.

Acknowledgements. We are grateful for support from DGAPA-UNAM and CONACyT (Mexican agencies).

References

1. Zipf, G.K.: Human Behavior and the Principle of Least-Effort. Addison-Wesley, Reading (1949)
2. Benford, F.: The Law of Anomalous Numbers. Proceedings of the American Philosophical Society 78(4), 551–572 (1938)
3. van der Galien, J.G. (November 08, 2008), See: http://en.wikipedia.org/wiki/Zipfs_law
4. Pietronero, L., Tosatti, E., Tosatti, V., Vespignani, A.: Physica A 293, 297 (2001)
5. García-Morales, V., Pellicer, J.: Physica A 361, 161 (2006)
6. CIA-The World Factbook, <http://numbrary.com/sources/f33de7d1aa-cia-the-world-factboo>
7. Leech, G., Rayson, P., Wilson, A.: Word Frequencies in Written and Spoken English: based on the British National Corpus. Longman, London (2001)
8. Barrueco, J.M.: Ranking of journals by impact factor. The Econometrics Journal, <http://www.feweb.vu.nl/econometriclinks/rankings/>
9. Naumis, G.G., Cocho, G.: Physica A 387, 84 (2008)
10. Beltrán Del Río, M., Cocho, G., Naumis, G.G.: Physica A 387, 5552 (2008)
11. Altamirano, C., Robledo, A.: (to be submitted)