Scientific Research Publishing

Advances in Pure Mathematics, 2023, 13, 174-180
https://www.scirp.org/journal/apm
ISSN Online: 2160-0384
ISSN Print: 2160-0368

Zipf's Law, Benford's Law, and Pareto Rule

Oded Kafri

Kafri Nihul Ltd., Tel Aviv, Israel
Email: kafri.entropy@gmail.com

How to cite this paper: Kafri, O. (2023) Zipf's Law, Benford's Law, and Pareto Rule. *Advances in Pure Mathematics*, **13**, 174-180.
https://doi.org/10.4236/apm.2023.133010

Received: February 23, 2023
Accepted: March 19, 2023
Published: March 22, 2023

Abstract

From a basic probabilistic argumentation, the Zipfian distribution and Benford's law are derived. It is argued that Zipf's law fits to calculate the rank probabilities of identical indistinguishable objects and that Benford's distribution fits to calculate the rank probabilities of distinguishable objects. *i.e.* in the distribution of words in long texts all the words in a given rank are identical, therefore, the rank distribution is Zipfian. In logarithmic tables, the objects with identical 1st digits are distinguishable as there are many different digits in the 2nd, 3rd… places, etc., and therefore the distribution is according to Benford's law. Pareto 20 - 80 rule is shown to be an outcome of Benford's distribution as when the number of ranks is about 10 the probability of 20% of the high probability ranks is equal to the probability of the rest of 80% low probability ranks. It is argued that all these distributions, including the central limit theorem, are outcomes of Planck's law and are the result of the quantization of energy. This argumentation may be considered a physical origin of probability.

Keywords

Zipf's Law, Benford's Law, Pareto 20 - 80 Rule, Planck's Law, Max Entropy

1. Introduction

Zipf's law and Benford's law are long-tail rank distributions appearing in many copious statistical ensembles [1]. Both laws are considered empirical laws. In 1881, Newcomb [2] found that the probability distribution $p(n)$ of the decimal digits in the 1st digits of the logarithmic table obeys $p(n)=\log\left(1+\frac{1}{n}\right)$, where $1\le\lfloor n\rfloor\le 9$. Benford [3] found in 1938 that Newcomb's distribution applies to many more ensembles and not only to the logarithmic table. Later [4] [5] the law generalized for N ranks to be,

$$p_B(n,N)=\frac{\ln\left(1+\frac{1}{n}\right)}{\ln(N+1)}. \tag{1}$$

Eleven years later, in 1949, Zipf [6] discovered that in long texts, in several languages, the most frequent word appears twice as much as the second most frequent word, the second most frequent word appears twice as much as the fourth frequent word, and so on. The Zipfian distribution, similarly to Benford's law, appears in many ensembles, like populations of cities, bestsellers lists, etc. Zipf's law can be written [7] [8] as,

$$p_z(n,N)=\frac{1}{nH_N}, \tag{2}$$

where $H_N=\sum_{n=1}^{N}\frac{1}{n}$ is the N^{th} harmonic number.

Both Zipf's law and Benford's law are obtained from the maximum entropy distribution of indistinguishable balls in N distinguishable boxes, where the boxes are the ranks and the number of the balls is much larger than the number of boxes [8]. In **Figure 1**, the Benford distribution and Zipf distribution for 10 ranks are plotted. It is seen that Benford's law and Zipf's law are similar but not identical.

Hereafter, we derive both laws using basic probabilistic tools and explain the differences between them. In addition, we derive the Pareto 20 - 80 rule of thumb for Benford's law and discuss their origin and limitations.

2. Zipf's Law

Suppose that there are N identical biscuits and a mouse in a closed space. The mouse eats every day one biscuit. What is the probability of a biscuit being eaten on the d day?

The maximum survival days n that a biscuit has at the day d is,

$$n=N+1-d,$$

where $1\le\lfloor d\rfloor\le N$.

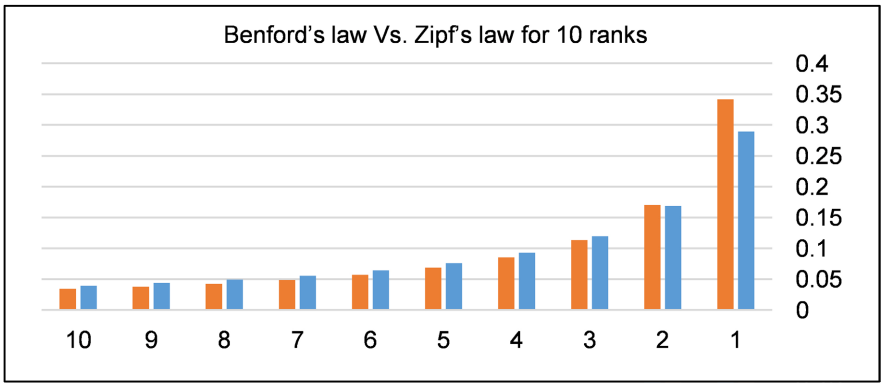

Figure 1. The red bars are the Zipfian distribution and the blue bars are Benford's law distribution for 10 ranks.

On the first day, $d=1$, the biscuit has maximum $n=N$ days to survive. Where $d=N$, the biscuit has only $n=1$ day. The probability p of the biscuit to be eaten is inversely proportional to n, namely, $p \propto 1/n$, therefore, the normalized probability distribution is,

$$p_Z(n,N)=\frac{\frac{1}{n}}{\sum_{n=1}^{N}\frac{1}{n}}=\frac{1}{nH_N},$$

which is Zipf's law.

We see that the probability of a biscuit being eaten on the day n obeys Zipf's law. This model, which is similar to the coupon collector problem, is identical to the word distribution of long texts. Suppose that one wants to write a text of N words. The first word has a probability of $\frac{1}{N}$, the second word $\frac{1}{N-1}$, etc. In the discussion, we explain why the Zipf distribution is so general.

3. Benford's Law

Benford's law is obtained by applying the Riemann sum to Zipf's law [8] [9]. If we assume that n is continuous, then,

$$\frac{1}{n}\approx\int_{n'=n}^{n+1}\frac{\mathrm{d}n'}{n'}=\ln\left(1+\frac{1}{n}\right) \quad \text{and} \quad H_N\approx\int_{n'=1}^{N+1}\frac{\mathrm{d}n'}{n'}=\ln(N+1)$$

Substitute these integrals in Zipf's law (Equation (2)) and we obtain Benford's law (Equation (1)).

Benford's law seems to approximate the more accurate Zipf's law. However, under certain conditions, Benford's law is more accurate than Zipf's law. For example, suppose that a pig that eats M biscuits per day replaces the mouse in the example above. In this case, a day becomes a rank that contains M biscuits. Since in a day there are M biscuits, the probability of a biscuit m to be eaten in the n day is,

$$p_Z(n,m,NM)=\frac{1}{H_{NM}}\cdot\frac{1}{nM+m}. \tag{3}$$

The probability to be eaten in the whole n^{th} day is

$$p_Z(n+1,NM)=\frac{1}{H_{NM}}\cdot\sum_{m=1}^{M+1}\frac{1}{nM+m}.$$

Since $\sum_{m=1}^{M+1}\frac{1}{nM+m}=H_{(n+1)M}-H_{nM}$, therefore,

$$p_Z(n+1,NM)=\frac{H_{(n+1)M}-H_{nM}}{H_{NM}}.$$

for $M\gg 1$, we can use the approximation

$$\lim_{M\to\infty}H_M=\ln(M)+\gamma,$$

where $\gamma\approx 0.577$ is the Euler-Mascheroni constant. Therefore,

$$p_Z\left(n+1,NM\right)\approx\frac{\ln\left[\left(n+1\right)M\right]-\ln\left[nM\right]+\gamma-\gamma}{H_{NM}}=\frac{1}{H_{NM}}\ln\left(1+\frac{1}{n}\right). \quad (4)$$

Equation (4), when renormalized, yields Benford's law. It is seen that Benford's law is obtained when there are sub-distributions inside Zipf's ranks.

4. Pareto 20 - 80 Rule of Thumb

In 1906, Italian economist Vilfredo Pareto [10] observed that 20% of the people in his country owned 80% of the nation's wealth. That rule was found to apply with uncanny accuracy to many situations and be useful in many disciplines, including the study of business productivity. Hereafter we show that the Pareto principle can be easily calculated from Benford's law. To do so we have to find the rank $\bar{n}$ which is the sum of the probabilities up to $\bar{n}$ is equal to the sum above it. In Benford's law, the rank $\bar{n}$ obeys,

$$\sum_{n=1}^{\bar{n}}\ln\left(1+\frac{1}{n}\right)=\sum_{n=\bar{n}}^{N+1}\ln\left(1+\frac{1}{n}\right),$$

which yields; $2\ln\left(\bar{n}+1\right)=\ln\left(N+1\right)$, or

$$\bar{n}=\sqrt{N+1}-1. \quad (5)$$

The Pareto ratio is simply,

$$\frac{\bar{n}}{N+1}:\frac{N+1-\bar{n}}{N+1} \quad (6)$$

Therefore $\bar{n}/\left(N+1\right)$ is the fraction of the ranks that have equal probability to the rest of the ranks and according to the Pareto rule is 0.2.

Zipf's law does not fit for Pareto ratio calculation as the distribution within the ranks does not exist and therefore none-integer $\bar{n}$ has no meaning. Benford's law is used for fraud detection of financial reports [11] [12]. However, Benford's distributions appear in many other statistics, of which a notable one is wealth distribution [13]. Pareto 20 - 80 distribution and Gini inequality index in free economies are in agreement with Benford's law [14]. However, as was shown Zipf's law, Benford's law and Pareto's rule are sensitive to the number of ranks N. Namely, the same distribution of probabilities yields different ratios between the ranks probabilities when N is changed. In Figure 2, we see that

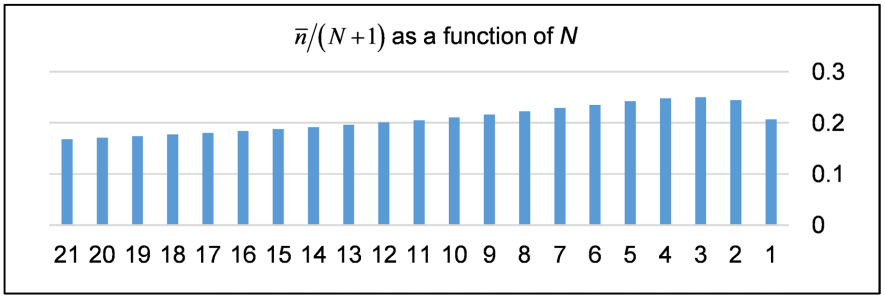

Figure 2. $\bar{n}/\left(N+1\right)$ is the fraction of the ranks that has the same probability as the rest of the ranks for the Benford distribution. The Pareto's 20 - 80 rule $\bar{n}/\left(N+1\right)\approx 0.2$ is valid in the vicinity of $N=10$ ranks.

around $N \approx 10$, the ratio 20 - 80 is a pretty good approximation of Benford's law distribution which fits better for the economy in which the incomes within the ranks are varying.

5. Discussion

The unequal probability distribution of the power laws is counterintuitive. If all the ranks have an equal probability to have an object, why they don't have an equal amount of objects? The explanation comes from statistical mechanics, An ensemble of ranks and their probabilities to have indistinguishable objects is analogous to a microcanonical ensemble of N boxes and $\langle n \rangle N$ balls, where $\langle n \rangle$ is the average number of balls in a rank. The thermodynamic microcanonical ensemble conserves material, volume, and energy. In the boxes and balls ensemble, the material is the boxes and their number N represents the conservation of volume. The number of balls represents the conservation of energy. According to the second law, in equilibrium, both the probabilities of the boxes to have a ball is equal *and*, all the microstates' probabilities are equal. A microstate (a state of the ensemble) is a distinguishable configuration of all the balls in all the boxes [7]. These requirements are an outcome of the second law, which one of its definitions states that in equilibrium the entropy is maximum. Planck calculated the distribution of the balls in the boxes in 1901 [15] [16]. He maximized the entropy of a set of distinguishable oscillators having an average energy $k_B T$, and each ball (photon) had an energy $h\nu$. Where k_B is the Boltzmann constant, T is the temperature, h is the Planck constant, and ν is the photon's frequency. The famous Planck result is,

$$n = \frac{1}{\exp\left(\frac{h\nu}{k_B T}\right) - 1}.$$

In the Planck equation, n is the occupation number of an oscillator in an ensemble in which the average energy is $k_B T$, and each photon has energy $h\nu$, therefore $\frac{k_B T}{h\nu}$ is the average number of photons in an oscillator. If we designate $\frac{k_B T}{h\nu} = \langle n \rangle$ we can write the Planck equation as,

$$[n] = \frac{1}{\exp\left(\frac{1}{\langle n \rangle}\right) - 1}. \tag{7}$$

In equilibrium for a given temperature and frequency all the oscillators should have the same number of photons $\langle n \rangle$. Since ν and T can have any value, $\langle n \rangle$ is not necessarily an integer, however, quantum mechanics enables, according to Equation (7), only an integer number of photons $[n]$ to exist. Therefore we can calculate the average number of balls $\langle n \rangle = f([n])$ as a function of the integer number of balls. In the case that $\langle n \rangle \gg 1$, $\exp(1/\langle n \rangle) \approx 1 + 1/\langle n \rangle$, we obtain that $n \approx \langle n \rangle$. This is the classical result in which the occupation number and the

number of balls are equal. Thus the probability is given by,

$$p(n)=\frac{1}{\langle n\rangle}\approx\frac{1}{n},$$

that when normalized to N boxes, yields Zipf's law as in Equation (2). In the general case Equation (7) yields

$$p(n)=\frac{1}{\langle n\rangle}=\ln\left(1+\frac{1}{n}\right). \tag{8}$$

When Equation (8) is normalized to N boxes it becomes Benford's law of Equation (1).

In the case when $\langle n\rangle \ll 1$, $\frac{1}{\langle n\rangle}\gg 1$, or $\exp\left(-\frac{1}{n}\right)-1\approx\exp\left(-\frac{1}{n}\right)$, the probability to find n balls, namely

$$p(n)=\frac{1}{\langle n\rangle}=\frac{1}{\exp\left(\frac{1}{n}\right)-1}\approx\exp\left(-\frac{1}{n}\right). \tag{9}$$

When normalized Equation (9) yields the canonical distribution namely.

$$p(n,N)=\frac{\exp\left(-\frac{1}{n}\right)}{\sum_{n=1}^{N+1}\exp\left(-\frac{1}{n}\right)}.$$

The normalization factor

$$Z=\sum_{n=1}^{N+1}\exp\left(-\frac{1}{n}\right)\approx\sum_{n=1}^{\infty}\exp\left(-\frac{1}{n}\right)$$

is the canonical partition function, which yields the central limit theorem in the limit of very small $\langle n\rangle$ [9].

6. Summary

Zipf's law, Benford's law, and Pareto's 20 - 80 rule are considered empirical laws. We argue that Zipf's law is the rank distribution of indistinguishable objects, while Benford's law is the rank distribution in which the objects within the rank are distinguishable. Pareto's 20 - 80 ratio, was found to be in good agreement with Benford's law in the vicinity of 10 ranks. It has also been argued that all these distributions, including the central limit theorem, can be derived from Planck's law and are the result of the quantization of energy. This argumentation may be considered a physical origin of probability.

Acknowledgements

I thank H. Kafri and E. Fishof for reading the manuscript and for their useful comments.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Tao, T. (2009) Benford's Law, Zipf's Law, and the Pareto Distribution. https://terrytao.wordpress.com/2009/07/03

[2] Newcomb, S. (1881) Note on the Frequency of Use of the Different Digits in Natural Numbers. *American Journal of Mathematics*, **4**, 39-40. https://doi.org/10.2307/2369148

[3] Benford, F. (1938) The Law of Anomalous Numbers. *Proceedings of the American Mathematical Society*, **78**, 551-572.

[4] Kafri, O. (2009) Entropy Principle in Direct Derivation of Benford's Law.

[5] Kafri, O. and Kafri, H. (2013) Entropy-God's Dice Game. CreateSpace, 208-209. http://www.entropy-book.com/

[6] Zipf, G.K. (1949) Human Behavior and the Principle of Least-Effort. Addison-Wesley.

[7] Powers, D.M.W. (1998) Applications and Explanations of Zipf's Law. *NeMLaP3/CoNLL*98: *ACL*, 151-160. https://doi.org/10.3115/1603899.1603924

[8] Kafri, O. (2020) Microcanonical Partition Function.

[9] Kafri, O. (2016) A Novel Approach to Probability. *Advances in Pure Mathematics*, **6**, 201-211. https://doi.org/10.4236/apm.2016.64017

[10] Pareto, V. (1964) "Cours d'Économie Politique": Nouvelle édition par G.-H. Bousquet et G. Busino. Librairie Droz, Geneva. https://doi.org/10.3917/droz.paret.1964.01

[11] Nigrini, M. (1996) A Taxpayer Compliance Application of Benford's Law. *The Journal of the American Taxation Association*, **18**, 72-91.

[12] Kossovsky, A.E. (2014) Benford's Law: Theory, The General Law of Relative Quantities, and Forensic Fraud Detection Applications., World Scientific Pub. Co. https://doi.org/10.1142/9089

[13] Kafri, O. (2018) Money Physics and Distributive Justice. CreateSpace, 94-95.

[14] Kafri, O. and Fishof, E. (2016) Economic Inequality as a Statistical Outcome. *Journal of Economics Bibliography*, **3**, 570-576.

[15] Planck, M. (1901) On the Law of Distribution of Energy in the Normal Spectrum. *Annalen der Physik*, **4**, 553. https://doi.org/10.1002/andp.19013090310

[16] Kafri, O. and Kafri, H. (2013) Entropy-God's Dice Game. CreateSpace, 198-201. http://www.entropy-book.com