Zipf's Law Everywhere

Article · January 2002

1 author:

Wentian Li
The Feinstein Institute for Medical Research
235 PUBLICATIONS   **10,566** CITATIONS

Some of the authors of this publication are also working on these related projects:

Probability distributions in complex phenomena View project

Multi-marker methods for the identification of novel asthma phenotypes View project

Zipf's Law Everywhere

Wentian Li[1]

Abstract. At the 100th anniversary of the birth of George Kingsley Zipf, one striking fact about the statistical regularity that bears his name, Zipf's law, is that it seems to appear everywhere. We may ask these questions related to the ubiquity of Zipf's law: Is there a rigorous test in fitting real data to Zipf's law? In how many forms does Zipf's law appear? In which fields are the data sets claiming to exhibit Zipf's law?

Keywords: Zipf´s law, ranking, language, population, internet, economics, bibliometrics, natural phenomena

1. Testing Zipf's law against alternative functions

Claiming a Zipf's law in a data set seems to be simple enough: if n values, x_i ($i = 1, 2, \ldots, n$), are ranked by $x_1 \geq x_2 \geq \ldots x_r \ldots \geq x_n$, Zipf's law states,

$$(1) \qquad x_{(r)} = \frac{C}{r^{\alpha}}$$

where the parameter value, α, is usually close to 1, implies that the $x_{(r)}$ versus r plot on a log-log scale will be a straight line with a negative slope α close to -1. If we assume $x_{(r)}$ as a random variable, from the statistical modeling point of view, Zipf's law is a model of the average of $x_{(r)}$ or $\log(x_{(r)})$ as a linear function (linear regression) of $\log(r)$ (with $c = \log(C)$):

$$(2) \qquad E(\log x_{(r)}) = c - \alpha \log(r).$$

However, visual inspection of the log-log plot of the ranked data is not a rigorous test. What if another functional form fits the same data better? Indeed, there are several functions that have been proposed as alternatives to Zipf's law in fitting the ranked data, such as (i) the Yule distribution (Yule 1925):

$$(3) \qquad x_{(r)} = \frac{C}{r^{\alpha} B^r}$$

or in the statistical modeling framework (with $c = \log(C)$, $b = \log(B)$):

---

[1] Adress correspondence to: Wentian Li, Center for Genomics and Human Genetics, North Shore LIJ Research Institute, 350 Community Drive, Manhasset, NY 11030, USA. E-mail: wli@nslij-genetics.org

(4) $E(\log x_{(r)}) = c - \alpha \log(r) - be^{\log(r)}$;

(ii) a variant of the log-normal distribution:

(5) $E(\log x_{(r)}) = c - \alpha \log(r) - b(\log(r))^2$;

or, (iii) a variant of the Weibull distribution:

(6) $E(\log x_{(r)}) = c - \alpha \log(r) - be^{\beta \log(r)}$;

where $0 < \beta < 1$.

In all three examples of an alternative function, there is a systematic modulation of the basic power-law structure in Zipf's law. In fact, such systematic deviation from the straight line in log-log plot is indeed present in some claimed Zipf's law patterns (Piqueira et al. 1999), which naturally causes a legitimate concern that some other claimed Zipf's laws in the literature may not be really Zipf's law.

A similar caution was raised in the example of the claimed Zipf's law pattern in DNA oligonucleotide frequencies (Mantegna et al. 1944}. There were many criticisms of this work (see, e.g., Bonhoeffer et al. 1996; Israeloff, Kagalenko, and Chan 1996; Voss 1996; Li 1996). One of the specific criticisms is that the data could be fitted by an alternatively function, the Yule distribution (Martindale and Konopka 1996).

It should be pointed out that it is not enough to reject the Zipf's law only because another function fits the ranked data better. The alternative function should not have too many extra parameters in achieving the better fit. The topic of statistical model selection is extensively discussed in Burnham and Anderson (2002). It is conceivable that we may use either the Bayesian information criterion (BIC) (Schwarz 1976) or Akaike information criterion (AIC) (Akaike 1974; Parzen, Tanabe, and Kitawa 1998) in selecting Zipf's law among alternatives. Some related ideas were also discussed in Quandt (1964) and Urzua (2000).

2. Two forms of Zipf's law

Besides the familiar form of Zipf's law for ranked data, there is another equivalent form of Zipf's law (Miller 1965). Actually, the second form is the probability density function of $x_{(r)}$, $p(x)$. Considering this simple procedure: switch the rank r and ranked value $x_{(r)}$ axes, then reverse the direction of the $x_{(r)}$. The resulting plot is simply the accumulative distribution (not normalized) of $x_{(r)}$ (see, e.g., Urzua 2000; Rousseau 2002). In mathematical expression, it is:

(7) $$\frac{r(x)}{n} = 1 - \int_{\min(t)}^{x} p(t)dt.$$

Knowing $r(x)$, or equivalently, $x(r)$, the probability density function $p(x)$ can be obtained by

(8) $$p(x) = -\frac{1}{n}\frac{d}{dx}r(x) \quad \text{or,} \quad p(x) = -\left(n\frac{d}{dr}x(r)\right)^{-1}.$$

It can be easily shown that the Zipf's law in Eq. (1) is equivalent to the following form of the probability density function of $x_{(r)}$:

$$(9) \qquad p(x) = \frac{C^{1/\alpha}}{\alpha n} \frac{1}{x^{(1/\alpha)+1}} = \frac{A}{x^\beta},$$

with $A = C^{1/\alpha}/n\alpha$ and $\beta = \alpha^{-1} + 1$, Eq. (9) is also a power-law function. The exponent $\alpha = 1$ as proposed originally in Zipf's law leads to $\beta = 2$. Some of the claimed Zipf's law was indeed illustrated as a probability density function (Axtell 2001).

3. Phenomena claiming a Zipf's law pattern

3.1. Word usage in human languages

The variable x is the number of times a word is used in written human languages (Zipf 1932, 1949; Kucera and Francis 1967). The frequency of usage can also be extended to spoken languages (Dahl 1979), non-English or non-Latin languages (Rousseau and Zhang 1992), combination of words (Egghe 2000), etc. Many articles in this volume are devoted to reviews on this example (Rousseau 2002; Altmann 2002; Hřebíček 2002; Montemurro and Zanette 2002).

3.2. City populations

The variable x is the number of people living in a city (Zipf 1949; Hill 1970; Ijiri and Simon 1977; Rosen and Resnick 1980; Gabaix 1999; Knudsen 2001; Soo 2002; Brakman, Garretsen, and Marrewijk 2001). The Zipf's law pattern can be easily checked by obtaining large city population data from a World Almanac, as was done in (Gell-Mann 1994). The city population can also be extended to those of metropolitan area, tribal society, regional areas (Davis and Weinstein 2001), etc. In a recent most extensive analysis of city population in different countries, the exact form of Zipf's law (i.e. $\alpha = 1$) was confirmed in 20 out of 73 countries (Soo 2002).

3.3. Webpage visits and other internet traffic data

In 1997, as a webmaster for a human genetics resource site (http://linkage.rockefeller.edu/), I was curious about whether the number of website visits per month followed the Zipf's law pattern. A quick plot showed it did. Being excited, I wanted to check whether someone else had come up with the same idea before I started to write this up in a publication. My web search ended up at the computer science department of Boston University where the same Zipf's law pattern for webpage visits was already discovered (Cunha, Bestavros, and Crovella 1995)! In the last few years, the study of scaling behaviors in internet traffic (with Zipf's law included) has become one of the hottest topics in applied computer science (Glassman 1994; Crovella and Bestavros 1997; Barford et al. 1999; Huberman et al. 1998; Barabasi and Albert 1999; Breslau et al. 1999; Adamic and Huberman 2002; Mitzenmacher 2003).

3.4. Company sizes and other economic data

This is another example of an easily obtainable data from the World Almanac. A company can be ranked by the number of employees, revenue, profit, market cap, as well as many other measurements. Such ranking can also be done within certain industry or certain geographical locations. The income distribution (Aitchison and Brown 1954; Samuelson 1952; Aoyama et al. 2000) is famously related to Pareto's law (Pareto 1896), which is frequently indistinguishable from the Zipf's law (the only difference being whether the α value is equal to 1 or not). One of the recent large-scale analyses of US company sizes is presented in (Axtell 2001). A constant debate on economic data is whether these are distributed as power-law (e.g. Pareto, Zipf) or as log-normal (Aitchison and Brown 1954; Champernowne 1953; Axtell 2001; Mitzenmacher 2003), or perhaps other distributions (Dagum 1984; Dragulescu and Yakovenko 2001a,b; Azzalini and Kotz 2002}.

3.5. Science citation and other bibliometric data

Similar to the popularity of webpages, popularity of scientific papers can be measured by how many times it is cited by other scientists. Scientists can also be ranked by how many papers he/she publishes (a measure of "productivity"). Other "bibliometric" data include the frequency of library items being loaned/borrowed. A pioneer of bibliometric data analysis was Alfred Lotka (1926). The following papers can be consulted for more details on bibliometric analysis: (Fairthorne 1969; Wyllys 1981; White and McCain 1989; Hertzel 1987; Egghe 1991; Egghe and Rousseau 1990; Osareh 1996a,b; Silagadze 1997; Redner 1998}.

3.6. Scaling in natural and physical phenomena

Since it has been shown that an inverse power-law with exponent α in the ranked data is equivalent to an inverse power-law in the probability density function with the exponent β = (1/α) +1, and Zipf's law with α =1 corresponds to β = 2, we can bring many more observed scaling behavior (i.e. power-law behavior) (Schroeder 1991) as examples of Zipf's law.

For example, the famous Gutenberg-Richter law states that the number of earthquakes whose magnitude are larger the *M* is an exponential function of *M* (Sornette et al. 1996):

$$(10) \quad N(x > M) \propto e^{-bM} \quad \text{with } b \approx 1.$$

Note that Eq.(10) is an accumulative distribution of the probability density function, and earthquake magnitude is a logarithm of the energy released $M \propto \log(E)$. It can be shown that the probability density function for earthquake energy according to Gutenberg-Richter law is $p(E) \propto 1/E^{b+1} = 1/E^2$, same as would be predicted by the Zipf's law.

3.7. Not all data exhibit Zipf's law

Although the title of this article is "Zipf's law everywhere", it is, of course, not literally everywhere. We have already shown examples where systematic deviation is present in the log-log plot of the ranked data (Piqueira et al. 1999; Mantegna et al. 1994). Also, when the size of the data (*n*) is small, it is usually hard to be convincing that we observe a power-law function. For example, the usage of 20 amino acids in protein sequence does not follow Zipf's law (Gamow and Ycas 1955). The 26 letters also do not follow Zipf's law in an English text.

If the *x* variable is a derived quantity (as versus a direct observable), the exponent α depends on how *x* is derived. For example, in a study to rank genes in their ability to classify

cancer subtypes (Li and Yang 2002), the (log) likelihood under a statistical discriminant model is used. If this likelihood is normalized by the number of samples in the microarray experiment, the exponent α in the Zipf's plot will be altered. On the other hand, if a more direct measurement is used, it is possible to have a traditional Zipf's law (Furusawa and Kaneko 2002).

4. Conclusions

It is tempting to propose a universal mechanism for Zipf's law because of the impression that Zipf's law is everywhere. Indeed, very general mechanisms were proposed (Yule 1925; Simon 1955), which without doubt would explain a large number of observed Zipf's law patterns (for a review of the explanations of Zipf's law, see, e.g., (Mitzenmacher 2003)).

But is our impression correct? Some of the true Zipf's laws may not be even well known to be a Zipf's law because the data is not presented as a ranked data. As we know the second form of the Zipf's law, we should look for any probability density function of the form $1/x^2$. On the opposite end, many claimed Zipf's law patterns may not be true of Zipf's law after all. Some data might be fitted better by alternative functional forms which nevertheless were not looked into by researchers.

The lesson is that we should pay attention to the data first. We may re-discover new dataset which exhibit Zipf's law, and at the same time, reject some claims of the Zipf's law in the literature. Despite my best efforts to collect all claimed Zipf's law in a webpage (http://linkage.rockefeller.edu/wli/zipf/), such efforts seem to be less than perfect, and there are always false claims and missing ones.

Acknowledgements

References

Adamic, L.A., Huberman, B.A. (2002). Zipf's law and the internet. *Glottometrics 3, 143-150.*

Aitchison, J., Brown, J.A.C. (1954). On criteria for descriptions of income distribution. *Metroeconomica 6, 88-98.*

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19, 716-723.*

Altmann, G. (2002). Zipfian linguistics. *Glottometrics 3, 19-26.*

Aoyama, H., Souma, W., Nagahara, Y., Okazaki, M.P., Takayasu, H., Takayasu, M. (2000). Pareto's law for income of individuals and debt of bankrupt companies. *Fractals 8, 293-300.*

Axtell, R.L. (2001). Zipf distribution of U.S. firm sizes. *Science 293, 1818-1820.*

Azzalini, A., Kotz, S. (2002). *Log-skew-normal and log-skew-t distributions as models for family income data.* University of Padua Department of Statistical Sciences, preprint.

Barford, P., Bestavros, A., Bradley, A., Crovella, M. (1999). Changes in web client access patterns: characteristics and caching implications. *World Wide Web 2, 15-28.*

Barabasi, A.L., Albert, R. (1999). Emergence of scaling in random networks". *Science 286, 509-512.*

Bonhoeffer, S., Herz, A.V.M., Boerlijst, M.C., Nee, S., Nowak, M.A. May, R.M. (1996), Explaining 'linguistic features' of noncoding DNA. *Science 271(5245), 14-15.*

Brakman, S., Garretsen, H., Marrewijk, C. van (2001). *An Introduction to Geographical Economics*. Cambridge, England: Cambridge University Press.

Breslau, L., Cao, P., Fan, L., Philips, G., Shenker, S. (1999). Web caching and Zipf-like distributions: evidence and implications. *Proceedings of IEEE Infocom'99, 126-134*.

Burnham, K.P., Anderson, D.R. (2002²). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Berlin: Springer-Verlag.

Champernowne, D. (1953). A model of income distribution. *Economic Journal 63, 318-351*.

Crovella, M., Bestavros, A. (1997). Self-similarity in world wide web traffic: evidence and possible causes. *IEEE/ACM Transactions on Networking 5, 835-846*.

Cunha, C.R., A Bestavros, A., Crovella, M.E. (1995). Characteristics of WWW client-based traces. Boston University Computer Science Department, *Technical report TR-95-010.*

Dagum, C. (1984). Income distributions models. In: Kotz, S. Johnson, N.L., Read, C.B. (eds*.), Encyclopedia of Statistical Sciences, vol. IV: 21-34*. New York: Wiley.

Dahl, G. (1979). *Word Frequencies of Spoken American English*. Essex, CT: Verbatim.

Davis, D.R.., Weinstein, D.E. (2001). Bones, bombs and break points: the geography of economic activity. National Bureau of Economic Research, working paper 8517.

Dragulescu, A., Yakovenko, V.M. (2001a). Evidence for the exponential distribution of income in the USA. *The European Physical Journal B, 20, 585-589*.

Dragulescu, A., Yakovenko, V.M. (2001b). Exponential and power-law probability distributions of wealth and income in the United Kingdom and the United States. *Physica A, 299, 213-221*.

Egghe, L. (1991). The exact place of Zipf's and Pareto's law amongst the classical informetric laws. *Scientometrics 20, 93-106*.

Egghe, L. (2000). The distribution of N-grams. *Scientometrics 47, 237-252*.

Egghe, L., Rousseau, R. (1990). *Introduction to Informetrics: Quantitative Methods in Library, Documentation and Information Science*. Amsterdam-New York: Elsevier.

Fairthorne, R.A. (1969). Empirical hyperbolic distributions (Bradford Zipf Mandelbrot) for bibliometric description and prediction. *Journal of Documentation 25, 319-343*.

Furusawa, C., Kaneko, K. (2003). Zipf's law in gene expression. *Physical Review Letters 90,088102.*

Gabaix, X. (1999). Zipf's law for cities: an explanation. *Quarterly Journal of Economics, 114, 739-767*.

Gamow, G., Ycas, M. (1955). Statistical correlation of protein and ribonucleic acid composition. *Proceedings of the National Academy of Sciences 41(12), 1011-1019*.

Gell-Mann, M. (1994). *The Quark and the Jaguar*. New York: Freeman.

Glassman, S. (1994). A caching relay for the world wide web. *Computer Networks and ISDN Systems 27(2), 165-173*.

Hertzel, D.H. (1987). Bibliometrics, history of the development of ideas. In: *Encyclopedia of Library and Information Science vol 42, suppl 7, 144-211*. New York: Dekker.

Hill, B.M. (1970). Zipf's law and prior distributions for the composition of a population. *Journal of the American Statistical Association 65, 1220-1232*.

Hřebíček, L. (2002). Zipf's law and text. *Glottometrics 3, 27-38*.

Huberman, B.H., Pirollo, P.L.T., Pitkow, J.E., Lukose, R.M. (1998). Strong regularities in world wide web surfing. *Science 280, 95-97*.

Ijiri, Y., Simon, H.A. (1977). *Skew Distributions and the Sizes of Firms*. Amsterdam: North-Holland.

Israeloff, N.E., Kagalenko, M., Chan, K. (1996). Can Zipf distinguish language from noise in noncoding DNA? *(letters), Physical Review Letters 76(11), 1976*.

Knudsen, T. (2001). Zipf's law for cities and beyond – the case of Denmark. American *Journal of Economics and Sociology 60, 123-146.*

Kucera, H., Francis, W.N. (1967). *Computational Analysis of Present-Day American English.* Providence, RI: Brown University Press.

Li, W. (1996). Comments on 'Bell curves and monkey languages' (letters). *Complexity, 1(6), 6.*

Li, W., Yang, Y. (2002). Zipf's law in importance of genes for cancer classification using microarray data. *Journal of Theoretical Biology 219(4), 539-551.*

Lotka, A.J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences 16, 317-323.*

Mantegna, R.N., Buldyrev, S.V., AL Goldberger, A.L., Havlin, S., Peng, C.P., Simon, M., Stanley, H.E. (1994). Linguistic features of noncoding DNA sequences. *Physical Review Letters, 73, 3169-3172.*

Martindale, C., Konopka, A.K. (1996). Oligonucleotide frequencies in DNA follow a Yule distribution. *Computers & Chemistry 20, 35-38.*

Miller, G.A. (1965). Introduction. In: *The Psycho-Biology of Language.* Cambridge, MA: MIT Press.

Mitzenmacher, M. (2003). A brief history of generative models for power law and lognormal distributions. Harvard University EECS Department, preprint.

Montemurro, M.A., Zanette, D.H. (2002). New perspectives on Zipf's law in linguistics: from single texts to large corpora. *Glottometrics 4, 87-99.*

Osareh, F. (1996). Bibliometrics, citation analysis and co-citation analysis: a review of literature I. *Libri 46, 149-158;* II: *Libri 46, 217-225.*

Pareto, V. (1896). *Cours d'Economie Politique.* Geneva: Droz.

Parzen, E., Tanabe, K., Kitagawa, G. (1998). *Selected Papers of Hirotugu Akaike.* Berlin: Springer.

Piqueira, J.R., Monteiro, L.H., Magalhaes, T.M.de, Ramos, R.T., Sassi, R.B., Cruz, E.G. (1999). Zipf's law organizes a psychiatric ward. *Journal of Theoretical Biology 198, 439-443.*

Quandt, R.E. (1964). Statistical discrimination among alternative hypotheses and some economic regularities. *Journal of Regional Science 5, 1-23.*

Redner, S. (1998). How popular is your paper? An empirical study of the citation distribution. *European Physical Journal B 4,131-134.*

Rosen, K.T., Resnick, M. (1980). The size distribution of cities: an examination of the Pareto law and primacy. *Journal of Urban Economics 8, 165-186.*

Rousseau, R., Zhang, Q. (1992). Zipf's data on the frequency of Chinese words revisited. *Scientometrics 24(2), 201-220.*

Rousseau, R. (2002). George Kingsley Zipf: life, ideas, his law and informetrics. *Glottometrics 3, 11-18.*

Samuelson, P.A. (1992). Graduated income taxation, which reduces inequality, leaves Pareto's coefficient invariant: a pseudo-paradox that debunks Pareto's coefficient. *Journal of Economic Perspectives 6, 205-206.*

Schroeder, M. (1991). *Fractals, Chaos, Power Laws.* New York: Freeman.

Schwarz, G. (1976). Estimating the dimension of a model. *Annals of Statistics 6, 461-464.*

Silagadze, Z.K. (1997). Citations and the Zipf-Mandelbrot's law. *Complex Systems 11, 487-499.*

Simon, H.A. (1955). On a class of skew distribution functions. *Biometrika 42, 425-440.*

Soo, K.T. (2002). Zipf's law for cities: a cross country investigation. London School of Economics, preprint.

Sornette, D., Knopoff, L., Kagan, Y.Y., Vanneste, C. (1996). Rank-ordering statistics of extreme events: application to the distribution of large earthquakes. *Journal of Geophysical Research 101, 13883-13893.*

Urzua, C.M. (2000). A simple and efficient test for Zipf's law. *Economics Letters 66, 257-260.*

Venables, W.N., Ripley, B.D. (1999³). *Modern Applied Statistics with S-PLUS*. Berlin: Springer.

Voss, R.F. (1996). Linguistic features of noncoding DNA sequences - Comment" (letters), *Physical Review Letters 76(11), 1978.*

White, H., McCain, K.W. (1989). Bibliometrics. *Annual Review of Information Science Technology* 24,119-186.

Wyllys, R.E. (1981). Empirical and theoretical bases of Zipf's law. *Library Trends 30, 53-64.*

Yule, G.U. (1925). A mathematical theory of evolution based on the conclusions of Dr. J.C. Willis, F.R.S. *Philosophical Transactions B 213, 21-87.*

Zipf, G.K. (1932). Selected Studies of the Principle of Relative Frequency in Language. Cambridge, MA: Harvard University Press.

Zipf, G.K. (1949). *Human Behavior and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley.