Entropy related remarks on Zipf's law for word frequencies

J.M.M.J. Vogels

Abstract.

Zipf's law applies to word frequencies in linguistics, to the population of cities, to DNA base pair sequences and a number of other situations. In a variety of problems where a large number of objects is distributed over an infinite sequence of clusters, a relation between rank k of each cluster and its population n is found according to n~1/k. Using a multinomial distribution we argue that this relation is the limit case of an asymptotically increasing entropy. Zipf's law may therefore be based on a simple argument that comes to the maximization of the entropy under the restriction that n describes a summable series for k→ ∞.

keywords: statistical linguistics, Zipf, hyperbolic distribution, entropy, population of cities, DNA

Introduction

It was demonstrated by Zipf [1]-[3] that the frequencies of words used in language follow a remarkable regularity. If one arranges the words into a sequence with decreasing frequency n, the product of n and the rank number k (k=1, 2, 3, …) is nearly constant: n~1/k. In present day English a good accuracy has been established [4], as well as in Dutch [5]. The same regularity applies to the number n of inhabitants of the rank of large cities in several countries [6], to the frequencies n of noncoding DNA base pair sequences [7] and to a wide variety of other phenomena [2]. A generalized formulation is $n\sim 1/k^{\alpha}$ with $\alpha\neq 1$, but in most cases the value of α is not much different from 1. In a random situation where many objects are clustering into a infinite rank of sets, the number n in each set tends to follow a regularity with the rank number k according to n~1/k. Fedorowicz [8] traced several theoretical approaches to the problem. Some authors [2], [9] develop more or less detailed models on the special features of the problem under consideration. Well known is the model for word frequencies by Mandelbrot [9, pp. 239-244]. He assumes a lexicographical tree for the hierarchical process of choosing words. By generating nonsense words Li [10] finds the formation of a Zipf distribution. He concludes that in linguistics the Zipf-like law is purely due to the choice of the rank as the independent variable and that the law in natural language is not of a profound character.

The main problem with ad hoc reasoning on special situations is that the law exhibits itself in a number of phenomena that are not related to each other in their special mechanisms or features. Therefore one would expect a more general underlying principle, in spite of special features. Because we are dealing with the distribution of many objects over a rank of clusters, one may expect to find some entropy related rule: a Zipf-like law should have a large probability and therefore be general for various cases of stochastic distribution. Simon [11] derives for the linguistical problem a distribution $n\sim 1/k^{\alpha}$ on the basis of a beta function and assuming that for large rank numbers the probability of finding a new word is constant. For the population problem Hill [12] proposes an allocation of people according to the Bose-Einstein form, which is the most critical assumption in his derivations. Besides this the sophisticated level of his mathematical derivations suggests a high degree of complication, which in our opinion rather contrasts with the generality and simple character of Zipf's law. The same comment may be given on

Sichel's [13] derivation of a word frequency law and his introduction of a new family of compound Poisson distributions. Chen [14] presents a generalization of the work by Hill [12], using the multinomial distribution as a basis. The latter starting point seems to be the most universal one. In our opinion however on the same general basis a rather simple reasoning can be developed which leads to Zipf's law with much less effort.

Theory

For reasons of simplicity we discuss the problem of the distribution of words in language, analogous to the distribution of balls over a number of urns, which problem may easily be translated into other Zipf-like problems. The random distribution of N balls over a number K of equal urns obeys a multinomial probability function (see [15])

$$P = \frac{N! / K^N}{n_1!\, n_2! \ldots n_K!} \qquad (1)$$

where $N = n_1 + n_2 + \ldots + n_K$. We simplify P to a logarithm ln P with the Stirling formula, valid for large n:

$$\ln n! \approx n \ln n - n \,, \qquad (2)$$

which yields

$$\ln P = a - \sum_{k=1}^{K} (n_k \ln n_k) \; ; \; a = N (\ln N - \ln K) \,. \qquad (3)$$

With the use of the Lagrange multiplier method in n_k - space (see appendix) it is proven that the most probable distribution is a flat one, the uniform distribution

$$n_1 = n_2 = \ldots = n_K \,. \qquad (4)$$

This results from maximizing ln P under the condition that the total number N is given. The uniform distribution implies that the maximum of equipartition between the urns has the largest probability, as is well known.

In our problem of word frequencies however the distribution cannot be uniform, because it has to be summable for $k \to \infty$. If we represent the distribution by the function n(k) , we conclude that in the tail, asymptotically for $k \to \infty$, the function should converge more rapidly than according to

$$n(k) = c/k \; ; \; c = \text{constant}. \qquad (5)$$

This is the limit of the weakest summable convergence. Now one may suspect intuitively that the maximum of equipartition, and therefore of ln P , should be reached for the weakest allowed convergence of n(k). On a log-log scale the summability limit of (5) corresponds with a slope of $-\pi/4$ rad :
$\ln n = \ln c - \ln k$.

So at this point we have the problem that maximizing ln P in (3), under the condition that N is given, yields a uniform distribution: the scattering of balls on a half infinite interval. As stated already, we are dealing with clustering and therefore with a summable convergence of n(k). So the summability should be brought into the Lagrange method by means of an additional summability condition.

Summability

The requirement that n_k should be a summable series when we extrapolate for $k \to \infty$ can be expressed by the criterion

$$\lim_{k\to\infty} (\ln n_k) / (\ln k) < -1 . \qquad (6)$$

This statement is easily proven by writing $(\ln n_k) / (\ln k) \leq -1-\varepsilon$, for positive ε and sufficiently large k. The criterion discriminates on α in $n \sim 1/k^{\alpha}$, when $\alpha \neq 1$. The case of $n_k \sim 1/k$ yields a divergent sum, while $n_k \sim 1/k.(\ln k)^{\alpha}$ with $\alpha > 1$ converges. Therefore a criterion with $(\ln n_k) / (\ln k) \to -1$ would yield an undefinite situation. For the moment we neglect this finesse. We will return to this subject in the discussion.

From (6) we derive

$$\lim_{k\to\infty} \frac{-n_k \ln n_k}{n_k \ln k} > 1 \qquad (7)$$

so that for sufficiently large k now $- n_k \ln n_k > n_k \ln k$.
Then no summable entropy (see (3)) will exist unless the expression

$$A = \sum_k n_k \ln k \qquad (8)$$

constitutes a finite number. Now the equation (8), with a finite A, expresses the requirement of summability we sought for.

Results

Now we maximize ln P in (3) under the condition that N is limited, combined with the summability condition (8); see appendix. The resulting class of distributions obeys the power law

$$n(k) = \varepsilon N / k^{(1+\varepsilon)} ; \varepsilon > 0 . \qquad (9)$$

We conclude that a power law according to (9) maximizes the entropy. On a log-log scale the power law corresponds with a fixed slope. More rapid schemes of convergence, like an exponential one, would have an increasing slope downwardly as k→ ∞ and therefore would allow less equipartition.

Next we will demonstrate that the entropy expression ln P , as described by (3), increases as in (9) the exponent ε decreases towards 0. First we note that (9) can be integrated:

$$\int_1^{\infty} n(k)\, dk = N . \qquad (10)$$

Replacing also in (3) the finite sum by the integral

$$\ln P = a - \int_1^{\infty} n \ln n \, dk \qquad (11)$$

and substituting (9) we obtain after a straightforward integration:

$$\ln P = N (1 + 1/\varepsilon - \ln \varepsilon - \ln K) . \qquad (12)$$

We observe a uniform and even asymptotic increase of ln P as ε→0.

Discussion

A probability will never be larger than unity. Because always $P \leq 1$, not every positive value of ε is allowed in (12). The use of the Stirling approach (2) is only justified for large values of n. The weaker however the convergence of n is for k→ ∞ , the more influence there is from a long tail with small values of n. So weak convergence introduces large errors in the resulting sum (12). Then entropy is a bad measure for probability. The replacement of a sum by an integral can be expected to yield an error of minor importance. But after all the strong and uniform decrease of ln P with increasing ε is evident: probability increases as n(k) converges in a weaker way.

It is needless to say that a distribution of the type

$$n = c / (k_0 + k) \qquad (13)$$

has the same asymptotic behavior for $k >> k_0$ as (5). Zipf's law in a more general formulation is often written as [9]

$$n = c / (k_0 + k)^{(1+\varepsilon)} \qquad (14)$$

with, in a number of cases, a value $0 < \varepsilon < 1$ and not $\varepsilon = 0$. To some types of problems a more detailed analysis than our own applies and therefore our entropy consideration will not always be conclusive. But even then we see in (12) that ln P decreases strongly as ε increases so that large values of ε are improbable just for reasons of entropy. The less we know about special mechanisms, the more likely it is that we observe an asymptotic behavior with $\varepsilon << 1$.

In our derivation we introduced the summability criterion (6), which generated a power law (9). Probably instead of (9) also other classes of weakly converging functions may occur. We saw already a function $n \sim 1/k\ (\ln k)^{\alpha}$ with $\alpha > 1$. The criterion (6) is not conclusive for functions n(k) with logarithmic deviations from $n \sim 1/k$, as we see. Is it possible to formulate a fully conclusive criterion? Alas it is not. For every class of summable functions it is possible to find another class with a weaker convergence. Correspondingly for every criterion one may formulate another one which is more sensitive. We give the proof of this statement in the table.

n(k)	$\int^{k} n\, dk \sim$	**criterion:** $\lim\limits_{k\to\infty}$ **u < -1 with**
$1/k^{\alpha}$	$1/k^{\alpha-1}$	u = ln n / ln k ; see (6)
$1/k.(\ln k)^{\alpha}$	$1/(\ln k)^{\alpha-1}$	u = ln (n.k) / ln (ln k)
$1/k.(\ln k).(\ln(\ln k))^{\alpha}$	$1/(\ln(\ln k))^{\alpha-1}$	u = ln (n.k.ln k) / ln(ln(ln k))
……	……	……

Table of functions n(k) with increasing weakness of summability and corresponding summability criteria. Always $\alpha > 1$.

Extension of the table yields every desired weakness of the summability of n(k). But it is clear that all functions n(k) approach in their asymptotics the one of $n \sim 1/k$. It may be doubted whether any criterion sharper than (6) would be of practical interest, when viewed in the light of empirical data. Would one reach the accuracy to distinguish $\ln n = c - \ln k - \alpha \ln(\ln k)$ from $\ln n = c - \ln k$ (c constant)? Nevertheless our own argument with the power law (9) cannot be used to forbid other functions approaching $n \sim 1/k$, because a sharper summability criterion than (6) cannot be excluded.

Thus far our argument concerns the asymptotics for k→ ∞ . For the first few rank numbers still all kinds of schemes are open. A relation like (14), with an additional parameter k_0 , even contains a somewhat larger probability P for $\varepsilon << 1$ than the simple one of (9). Every additional parameter in a distribution however implies some structure in the problem. The more parameters there are, the more structure and the less stochastic behavior. Therefore a problem without any structure at all and with a maximum of entropy should contain no additional parameters. Fully unstructured situations should yield a simple $n \sim 1/k$ (see (5)). This relation has indeed been observed in the first few rank numbers of the distribution of population over large cities [6].

We conclude that indeed the weakest acceptable convergence of n(k) for $k \rightarrow \infty$ yields the most probable distribution function. Therefore a tail in the distribution according to (5) is simply the limit case of maximum probability under the restriction of summability. This result may be significant in many stochastic situations where a large number of objects is distributed over an infinite rank of clusters, provided that further restrictions are absent.

The presence of additional restrictions in the problem may have major consequences. In many kinds of problems a more rapidly converging distribution is found, like the (exponential) Boltzmann distribution in statistical mechanics. This however follows in the same way as (4) or (9), if in the Lagrange multiplier method the additional condition is imposed that the total amount of energy in the system is given (see the appendix):

$$\sum_{k=1}^{K} n_k E_k = \mathbf{E} . \qquad (15)$$

Our own argument concerns a situation without such a constraint. Additional conditions may strongly change the whole picture.

Conclusion

Our final conclusion is that the hyperbolic law $n(k) \sim 1/k$ has a general character, in spite of special mechanisms. Those may be different in several situations, without much influence on the resulting distribution. Zipf's law turns out to be simply the limit of maximum probability (entropy), in the case that a large number of objects is distributed over an infinite sequence of clusters.

Appendix

We use the Lagrange multiplier method [16] in the derivation of (4), of (9) and of the Boltzmann distribution (15).

Maximize ln P in (3) under the condition that $N = n_1 + \ldots + n_K$ is given. The gradients of ln P and N in n_k - space should be aligned (anti-)parallel. So $\partial \ln P / \partial n_k = \lambda \, \partial N / \partial n_k$. The result is that

$\ln n_k = -\lambda'$, or $n_1 = n_2 = \ldots = n_K$; see (4).

To derive the power law (9) we maximize ln P in (3) under the conditions that again N is given, but also that A in (8) has a finite value: $\partial \ln P / \partial n_k = \lambda \, \partial N / \partial n_k + \mu \, \partial A / \partial n_k$. Here $n_k = \lambda' \, k^{-\mu}$; see (9).

To obtain the Boltzmann distribution we impose as a second condition that $\mathbf{E} = n_1E_1 + n_2E_2 + \ldots + n_KE_K$ is given (15). Now $\partial \ln P / \partial n_k = \lambda \, \partial N / \partial n_k + \mu \, \partial \mathbf{E} / \partial n_k$, so that $\ln n_k = - \lambda' - \mu \, E_k$. With a proper choice of the values of λ' and μ this constitutes a Boltzmann distribution.

Acknowledgement

The author would like to thank dr. Rob van den Berg for drawing his attention to Zipf's distribution and for providing him with some references on the subject.

J.M.M.J. Vogels, Eindhoven, The Netherlands (2008).

References

[1] G.K. Zipf, 'Selective Studies of the Principle of Relative Frequency in Language', MIT Press, Cambridge MA (1932)

[2] - - - , 'Human Behavior and the Principle of Least Effort', Addisson-Wesley, Reading MA (1965)

[3] - - - , 'The Psycho-Biology of Language: An Introduction to Dynamic Philology', MIT Press, Cambridge MA (1965)

[4] H. Kučera and W. Nelson Francis, 'Computational Analysis of Present-Day American English', Brown University, Providence RI (1967)

[5] J.W. Nienhuys, private communication (1995)

[6] B. Berry and W. Garrison, Ann. Assoc. Amer. Geographers 48(1958)83

[7] R.N. Mantegna, S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.-K. Peng, M. Simons and H.E. Stanley, Phys. Rev. Lett. 73(1994)3169

[8] J. Fedorowicz, J. Amer. Soc. Inform. Science (1982)285

[9] B. Mandelbrot, 'Fractals: Form, Chance and Dimension', Freeman, San Francisco (1977)

[10] W. Li, IEEE Trans. Inform. Theory 38(1992)1842

[11] H.A. Simon, Biometrika 42(1955)425

[12] B.M. Hill, J. Amer. Statist. Assoc. 69(1974)1017

[13] H.S. Sichel, J. Amer. Statist. Assoc. 70(1975)542

[14] W.-C. Chen, J. Appl. Prob. 17(1980)611

[15] CRC Handbook of Probability and Statistics, 2 nd ed., The Chemical Rubber Co. (1968)

[16] R. Courant and D. Hilbert, 'Methoden der Mathematischen Physik I', 3 auflage, Springer Verlag, New York (1968) ISBN 0-387-04177 X