CEE Centro de Estudios Económicos
www.colmex.mx El Colegio de México, A.C.

Serie documentos de trabajo

A SIMPLE AND EFFICIENT TEST FOR ZIPF'S LAW

Carlos Urzúa

DOCUMENTO DE TRABAJO

Núm. VIII - 1999

A simple and efficient test for Zipf's law*

Carlos M. Urzúa

El Colegio de México† and ITESM-CCM

March 2, 1999

ABSTRACT

This paper presents a simple and locally optimal test for Zipf's law. Its use is illustrated in the case of the largest US metropolitan areas. An objection to the general relevance of that law is also presented.

Keywords: Zipf's law; rank-size law

JEL classification: C12; R12

*This research is part of a project on Geography and Economic Development sponsored by the Inter-American Development Bank at El Colegio de México.

†Camino al Ajusco 20, México, DF 01000, México. Tel: (52) 5449-3000; fax: (52) 5645-0464; email: curzua@colmex.mx.

I. INTRODUCTION

Consider the ordered sequence of n data values

$$x_{(1)} \geq x_{(2)} \geq ... \geq x_{(r)} \geq ... \geq x_{(n)} \tag{1}$$

where r is the rank, and $x_{(r)}$ is the size (like, for instance, the size of a city). Zipf's law, also known as the rank-size law, asserts that a graph of the rank against the size would render a perfect rectangular hyperbola; that is, for some constant c and all r,

$$r x_{(r)} = c \quad . \tag{2}$$

Since Zipf (1949), it has been customary to verify that law by simply plotting the natural logarithm of the rank against the log of the size in the hope of finding a straight line with a slope equal to minus one (as implied by Eq. (2)). More formally, a "test" is sometimes constructed by estimating, through ordinary least squares, the regression

$$\ln r = \beta_1 + \beta_2 \ln x_{(r)} + \varepsilon_r \tag{3}$$

to check if the estimate of β_2 is "close" to -1. Needless to say, this procedure is inefficient: Since r is an integer, the distribution of ε in Eq. (3) is far from being normal.

As it has been forcefully noted by several authors over the years (e.g., Quandt, 1964, Rapoport, 1978, and Kamecke, 1990), before testing for Zipf's law one has to make explicit the underlying probabilistic process that is behind it. To put it in a different way: one has to translate the rank-size relation in Eq. (2) into a size-frequency relation. For that end,

let $f(x)$ be the relative frequency corresponding to a set of n objects. Then, the rank of an object of size x is given by

$$R(x) = n \int_x^\infty f(z)dz, \tag{4}$$

where we have assumed, without loss of generality, that all the objects have different sizes, and that the probability density is continuous. As a consequence, Eq. (2) can be recast into a probabilistic framework as: $R(x) = c/x$. Taking the derivative with respect to x in this last equation and in Eq. (4), we can finally found that Zipf's law implicitly states that

$$f(x) = \frac{c}{n}(\frac{1}{x^2}). \tag{5}$$

Hence, all tests should focus on the particular power law stated in Eq. (5), or on its discrete version. A simple and efficient test along that line is presented in the next section.

II. A LOCALLY OPTIMAL TEST

The density given in Eq. (5) is a special case of Pareto's well-known law (Pareto, 1897):

$$f(x) = \frac{\alpha}{\mu}(\frac{x}{\mu})^{-(\alpha+1)}, \qquad x \geq \mu \tag{6}$$

where $\alpha \geq 0$ and $\mu > 0$. It should be noted that although this density has in principle two parameters, μ is usually fixed by statistical design. Thus, we shall assume from now on that μ is estimated before hand using $x_{(n)}$ in Eq. (1).

Since Zipf's law is obtained when $\alpha = 1$ in Eq. (6), a simple test for such a null hypothesis could be derived under the premise that Pareto's law holds. This approach is followed, among others, by Kamecke (1990). A second, more robust approach would be to consider several

potential distributions, not only Pareto's law, and to search for the one that provides the best fit. This is the strategy followed by Quandt (1964). Finally, in this paper we propose a third approach, more akin to the methodology currently in use in Econometrics: After selecting a suitable density that nests Eq. (6), we derive a locally optimal test for Zipf's law by means of the Lagrange multiplier (LM) test.

For that end, consider the following density, more general than Pareto's law but still algebraically simple:

$$f(x) = \frac{\alpha}{\sigma}(1 + \frac{x - \mu}{\sigma})^{-(\alpha+1)} \qquad x \geq \mu. \tag{7}$$

Following the terminology in Johnson and Kotz (1970), the distribution implied by Eq. (7) is a member of Burr's family of distributions. In particular, if $\sigma = \mu$, then Pareto's law is obtained. Furthermore, if $\sigma = 1$, then the so-called Pareto density of the second kind is found. This distribution, also introduced by Pareto a century ago, produced the best fits, on the whole, in Quandt's comprehensive study mentioned earlier.

Returning to the general expression in Eq. (7), note that the corresponding log-likelihood function for a random sample of size n is given by

$$\ell(\sigma, \alpha) = -(\alpha + 1)\sum_{i=1}^{n} \ln(1 + \frac{x_i - \mu}{\sigma}) + n \ln \alpha - n \ln \sigma,$$

where, as explained earlier, μ can be replaced by $x_{(n)}$. Under this framework, a test for Zipf's law can be derived by devising a test for the null hypothesis $H_0 : \sigma = \mu; \alpha = 1$. In particular, the corresponding Lagrange multiplier test, as reviewed by, say, Engle (1984), is equal to $d' \mathfrak{I}^{-1} d$, where d is the score and $\mathfrak{I}$ is the information matrix, both evaluated under the null (there are no nuisance parameters in our case).

After some simple algebra, the Lagrange multiplier test for Zipf's law is found to be:

$$LMZ = 4n\left[z_1^2 + 6z_1z_2 + 12z_2^2\right] \overset{a}{\sim} \chi_2^2 \quad , \tag{8}$$

where

$$z_1 \equiv 1 - \frac{1}{n}\sum_{i=1}^{n} \ln \frac{x_i}{x_{(n)}} \quad \text{and} \quad z_2 \equiv \frac{1}{2} - \frac{1}{n}\sum_{i=1}^{n} \frac{x_{(n)}}{x_i} \quad .$$

The form of this omnibus test is rather intuitive. As can be easily checked, the log of the Zipf variate x/μ follows an exponential distribution with mean equal to one, while its inverse follows a uniform with mean equal to one-half. Thus, z_1 and z_2 simply measure the discrepancies between the population and sample means for those two transformations (in contrast, the Zipf variate does not have a mean!).

Under the null, *LMZ* is asymptotically distributed as a chi-square with two degrees of freedom. Although some LM statistics do not behave well in the case of small and medium-size samples (see, e.g., Urzúa, 1996), Table 1 presents evidence that this is not our case. As can be appreciated there, the asymptotic critical values for the most common significance levels can be safely used when $n \geq 50$.

III. An application to Urban Economics

As an example of the use of *LMZ*, consider the US metropolitan areas that, in 1991, had a population of 250,000 or more inhabitants (US Bureau of the Census, 1993, Table 42). For this data set, 135 areas in total, Krugman (1996, p. 40) presents a graph of the implied rank-size relation, while Gabaix (1998) gives the results of a regression similar to Eq. (3). Both authors claim that Zipf's law holds almost perfectly in this case. To do a formal testing, we

note that $x_{(135)} = 252{,}000$ (the population of Charleston, WV), and proceed to calculate the formulae in Eq. (8). The resulting value for LMZ is 3.16. Thus, using Table 1, we cannot reject the hypothesis that $\alpha = 1$ at a significance level of 10%.

We now enlarge the sample to consider, as it is typically done in urban studies, all the US metropolitan areas with a population of at least 100,000 inhabitants (the smaller areas are listed in Appendix II of the same source). For this sample, $n = 251$, $x_{(251)} = 100,000$ (the population of Gadsden, AL), and, finally, $LMZ = 21.92$. Hence, we now reject Zipf's law at a significance level of the order of 0.01%.

How can we reconcile those two conflicting results? The answer is that, as first observed by Herdan (1960), α depends on the sample size (as is evident when we contrast equations (5) and (6) above). Thus, strictly speaking, Zipf's law cannot hold except for a certain sample size, if at all.

IV. References

Engle, R.F., 1984. Wald, Likelihood ratio, and Lagrange multiplier tests in Econometrics. In: Griliches, Z., Intriligator, M. (Eds.), Handbook of Econometrics, vol. 2. North Holland, Amsterdam, pp. 775-826.

Gabaix, X., 1998. Zipf's law for cities: an explanation. Forthcoming in Quarterly Journal of Economics.

Herdan, G., 1960. Type-token Mathematics. Mouton, The Hague.

Johnson, N.L., and S. Kotz, 1970. Distributions in Statistics: continuous univariate distributions, vol.1. Houghton-Mifflin, Boston.

Kamecke, U., 1990. Testing the rank size rule hypothesis with an efficient estimator. Journal of Urban Economics 27, 222-231.

Krugman, P., 1996. The self-organizing economy. Blackwell, Oxford.

Pareto, V., 1897. Cours d'économie politique. F. Rouge, Lausanne.

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

Rapoport, A., 1978. Rank-size relations. In: Kruskal, H., Tanur, J.M. (Eds.), International encyclopedia of Statistics, vol. 2. Free Press, New York, pp. 847-854.

Urzúa, C.M., 1996, On the correct use of omnibus tests for normality. Economic Letters 53, 247-251. Erratum 1997, vol. 54, p. 301.

U.S. Bureau of the Census, 1993. Statistical abstract of the United States. Bureau of the Census, Washington.

Zipf, G.K., 1949. Human behavior and the principle of least effort: an introduction to human ecology. Addison-Wesley, Cambridge.

Table 1

Significance points for *LMZ*

n	10	15	20	25	30	50	100	200	∞
Level									
5%	6.19	6.14	6.09	6.08	6.03	5.98	5.98	5.99	5.99
10%	4.38	4.41	4.43	4.45	4.46	4.49	4.56	4.58	4.61

Source: Own Monte Carlo simulations using the inversion method, and after 100,000 replications.