Submitted to ***Nature***

U.S. Firm Sizes are Zipf Distributed

Robert L. Axtell
Center on Social and Economic Dynamics, The Brookings Institution
1775 Massachusetts Avenue, NW, Washington, D.C. 20036

February 22, 2001

Firm sizes have typically been described by lognormal distributions, with Pareto, Yule and related distributions accurately capturing the upper (large size) tail. Utilizing data on the entire population of U.S. firms, including small businesses, we find that the Pareto distribution well describes *the entire firm size distributio*n. Furthermore, the exponent of this distribution is essentially unity, thus we have the special case of the Zipf distribution. Data on self-employment, not normally included in small firm data, are consistent with the Zipf characterization. These results are shown to be robust to alternative definitions of firm size.

The Zipf distribution is a discrete, one-parameter, univariate distribution that has been used to describe various physical and social phenomena that are highly skewed in character. For instance, the frequency of word usage in printed texts is Zipf-distributed—the so-called Estroup-Zipf law—meaning that the frequency with which a word is used is inversely proportional to its rank, where the most commonly occurring word has rank 1. Similarly the distribution of city sizes in industrial countries are often Zipf-distributed.

The distribution of firm sizes in industrial countries is well-known to be highly skewed, with large numbers of small firms coexisting with small numbers of large firms. The stability of this distribution over time makes it, along with the distribution of city sizes, perhaps the most robust statistical regularity in all the social sciences[1].

Beginning with Gibrat, there is an established tradition of describing the distribution of firm sizes in industrial countries by lognormal distributions. This distribution is a direct consequence of the 'law of proportional effect,' whereby the growth of firms is treated as a random process with growth rates being independent of firm size.[2] In general, lognormal distributions are right skewed, meaning they are asymmetric with much of the probability mass to the right of the modal (most common) value. In the present context this amounts to modal firm size being less than the median, which in turn is less than the mean.

The upper tail of the firm size distribution has often been described by the Pareto distribution, also known as a power law or scaling distribution.[3] For a Pareto-distributed random variable, *S*, the cumulative tail distribution function is usually written as

$$\Pr\left[S \geq s\right] = \left(\frac{s_0}{s}\right)^{\alpha}, \ s > s_0$$

where s_0 is the minimum size and α is a parameter. Typical analyses of data on the very largest industrial firms yields values for α in the range 1.1 - 1.2, although systematic departures from this distribution are also known.[4] The special case of $\alpha = 1$ is known as the Zipf distribution. It has somewhat unusual properties insofar as its moments do not exist.

Utilizing data on U.S. firms from Compustat, a larger dataset than previous studies, Stanley *et al*. [1995] report that the distribution of U.S. firm sizes is very closely approximated by a lognormal distribution. They find that their fitted distribution predicts many more large firms than actually exist—that is, are too few large firms, contradicting previous results. Unfortunately, these results cannot not be used to draw general conclusions about U.S. firms because they are based on data that are unrepresentative of the overall population of U.S. firms.

The Compustat data covers essentially all *publicly-traded* firms in the U.S. In 1997, for instance, such firms numbered just under 11,000, although only 8200 reported having employees. While well-representing large firms, Compustat does not include privately-held firms, many of which are quite large in size. Indeed, data from the U.S. Economic Census puts the total number of firms in the U.S. at about 5.5 million, including over 16,000 having more than 500 employees. Furthermore, firm size data in the Census has a qualitatively different character than the Compustat data. Census data displays monotonically increasing numbers of increasingly smaller firms. This shape is one that the lognormal cannot reproduce, but suggests that a power law or similar distribution may apply over the whole size range. To get a sense of how

[1] Ijiri and Simon [1977], p. 2.

[2] Also known as 'Gibrat's law'; see Sutton [1997] for a review.

[3] Related distributions include the Yule and zeta; see, for instance, Ijiri and Simon [1977], especially chapter 7, and Mandelbrot [1997].

[4] See Ijiri and Simon [1977], chapter 11.

different these two datasets are, Table 1 compares them.[5] The size bins shown are those used by the SBA.

Size class	***Census/SBA***	***Compustat***
0	719,978	2576
0 - 4	3,358,048	2699
5 - 9	1,006,897	149
10 - 19	593,696	251
20 - 99	487,491	1287
100 - 499	79,707	2123
500+	16,079	4267

Table 1: Number of firms in various size classes (by number of employees) in the U.S. c. 1997, compared across two data sources

The mean firm size in the Compustat data (all firms) is 4605 while for the Census data it is much less. Clearly the Compustat data is unrepresentative with respect to small firms.

The binning of the data in Table 1 is of limited use primarily because it lumps all large size categories together, but also because the bins are of different sizes. We have obtained a tabulation from Census in which bins (except the first) are of increasing size in powers of 3. This is shown in Table 2.

Size class	***Economic Census***
0	719,978
1	1,026,469
2 - 4	1,611,601
5 - 13	1,342,582
14 - 40	575,228
41 - 121	190,236
122 - 364	53,513
365 - 1093	14,903
1094 - 3280	4909
3281 - 9841	1657
9842 - 29,524	610
29,525 - 88,573	178
88,574 - 265,720	48
265,721 - 797,161	6
797,162 and larger	0

Table 2: Number of firms in various size classes (by number of employees) in the U.S. c. 1997

The first entry requires explanation. Data on firm sizes was gathered in March of 1997. Firms that had receipts during 1997 but no employees as of March are shown in the 0 category. Clearly, such firms should be in one of the other size classes, and so this data is censored.

Neglecting this 0 size class, we transform the remainder of the data by taking logs of both coordinates. OLS on this transformed data yields a slope of -1.008, with a standard error (SE) of 0.040; adjusted R^2 = 0.994. The data are shown, along with the best fit line, in Figure 1, a so-called Zipf plot.

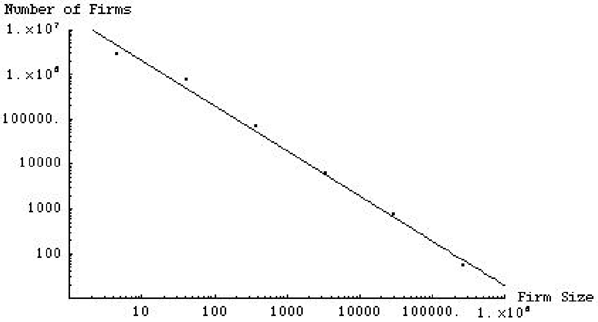

Figure 1: Distribution of U.S. firm sizes (by employees) for 1997, data combined from Census/SBA and Compustat

Interestingly, while there are some 4.8 million firms described in this figure (5.5 million - 700 thousand of size 0), there are another 15.5 million business entities in the U.S. that do not have any employees. These are predominantly individual proprietorships, and are reported in the Census data as 'non-employee' firms. These smallest of firms make up in number what they lack in size, accounting for nearly $600 billion in receipts in 1997. Perhaps somewhat surprisingly, if we include these firms in the overall firm size distribution we do not impair the quality of the Zipf distribution fit to the data, as shown in Figure 2.

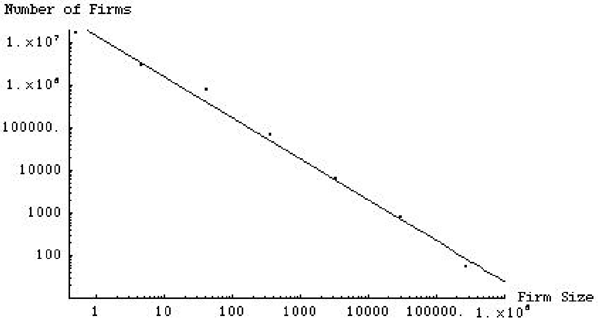

Figure 2: Distribution of U.S. firm sizes (by employees) for 1997, data combined from Census/SBA and Compustat together with self-employment data

Here OLS yields a slope of -0.963, SE = 0.038, and R^2 = 0.992.

So far the number of employees in a firm has constituted our measure of firm size. An interesting property of firm size distributions from previous studies is that the overall character of the distribution is independent of how size is defined. For example, in Ijiri and Simon [1977] size is based on revenue, while in

[5] Data from the U.S. Economic Census is based on Small Business Administration (SBA) tabulations.

other studies it is based on market capitalization. Here we check whether this situation yet holds, using data on receipts (revenue). Using again a specially prepared tabulation from Census, Figure 3 results.

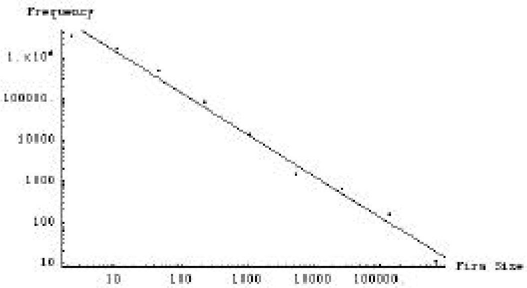

Figure 3: Distribution of U.S. firm sizes (by revenue, in $ million) for 1997, data from Census

The slope of this line is -1.039 with SE = 0.055 and adjusted $R^2 = 0.988$.

Just as the Zipf distribution of firm size is robust across varying definition of 'size', so too it is quantitatively invariant over time. We have repeated these calculations for data from 1992 and have obtained similar results.

There exists a theoretical tradition of describing Pareto and Zipf distributed size distributions with stochastic process models (e.g., Simon [1955] and Gabaix [1999]). The trouble with these explanations is that they are not written in terms of economic variables. As such, they are not microeconomic explanations. The only *microeconomic* model that gets these statistical features correct is described in Axtell [1999].

Acknowledgements

Dedicated to Herbert Simon, who initiated my interest in this subject in graduate school. Lectures on Zipf's law by Murray Gell-Mann [1990] at the Santa Fe Institute aided the development my thinking, as did the manuscript of Benôit Morel [1998]. Benôit Mandelbrot made several useful comments based on preliminary results. Thanks are also due Zoltan Acs for background information on small firms, and Carol Graham and Peyton Young for useful suggestions. Support from the National Science Foundation (grant 9820872) and the John D. and Catherine T. MacArthur Foundation to the Brookings Center on Social and Economic Dynamics is gratefully acknowledged. The author may be contacted via email at `raxtell@brookings.edu`.

References

Acs, Z., ed. 1999. *Are Small Firms Important? Their Role and Impact*. Kluwer Academic: Boston, Mass.

Acs, Z. and D. Audretsch. 1990. *Innovation and Small Firms*: MIT Press: Cambridge, Mass.

Axtell, R.L. 1999. The Emergence of Firms in a Population of Agents: Local Increasing Returns to Scale, Unstable Nash Equilibria, and Power Law Size Distributions. Working paper. The Brookings Institution. Available at www.brookings.edu/dynamics/ papers.

Champernowne, D.G. 1953. A Model of Income Distribution. *Economic Journal*, *63*, 318-351.

Dun and Bradstreet. 2001. *The Compustat User Guide*.

Gabaix, X. 1999. Zipf's Law of City Sizes: An Explanation. *Quarterly Journal of Economics*.

———. 2000. Zipf’s Law and the Growth of Cities. *American Economic Review*.

Gell-Mann, M. 1990. Zipf’s Law and Related Mysteries. Lectures given at the First Santa Fe Institute Winter School on Scaling and Complex Systems, Tuscon, Arizona.

———. 1994. *The Quark and the Jaguar*. W.H. Freeman: New York.

Gibrat, R. 1931. Les Inégalitiés Economiques: Applications aux inégalitiés des rechesses, à la concentration des entreprises, aux populations des villes, aux statistiques des familles, etc., d'une loi nouvelle, la loi de l'effet proportionnel. Librarie du Recueil Sirey: Paris.

Hart, P.E. and S.J. Prais. 1956. The Analysis of Business Concentration: A Statistical Approach. *Journal of Royal Statistical Society*, Series A, *119*, 150-181.

Ijiri, Y. and H. Simon. 1977. *Skew Distributions and the Sizes of Business Firms*. North-Holland: New York.

Mandelbrot, B. 1997. *Fractals and Scaling in Finance*. Springer-Verlag: New York.

Morel, B. 1998. . Working paper. Carnegie Mellon University: Pittsburgh, Penn.

Quandt, R. 1966. On the Size Distribution of Firms. *American Economic Review*, *56*: 416-432.

Simon, H. and C.P. Bonini. 1958. The Size Distribution of Business Firms. *American Economic Review*, *46*, 607-617.

Small Business Administration (www.sba.gov/advo/stats/ data.html)

Stanley, M.H.R., S.V. Buldyrev, S. Havlin, R.N. Mantegna, M.A. Salinger and H.E. Stanley. 1995. Zipf Plots and the Size Distribution of Firms. *Economics Letters*, *49*: 453-457.

Sutton, J. 1998. Gibrat's Legacy. *Journal of Economic Literature*, *35* (1): 40-59.

Zipf, G.K. 1949. *Human Behavior and the Principle of Least Effort*. Addision Wesley: Reading, Mass.