

## Zipf plots and the size distribution of firms

Michael H.R. Stanley<sup>a</sup>, Sergey V. Buldyrev<sup>a</sup>, Shlomo Havlin<sup>a, b</sup>,  
Rosario N. Mantegna<sup>a</sup>, Michael A. Salinger<sup>c, \*</sup>, H. Eugene Stanley<sup>a</sup>

<sup>a</sup>*Department of Physics, Boston University, 590 Commonwealth Avenue, Boston, MA 02215, USA*

<sup>b</sup>*Department of Physics, Bar-Ilan University, Ramat-Gan, Israel*

<sup>c</sup>*School of Management, Boston University, 704 Commonwealth Avenue, Boston, MA 02215, USA*

Received 14 November 1994; revised version received 9 March 1995; accepted 13 March 1995

---

### Abstract

We use a Zipf plot to demonstrate that the upper tail of the size distribution of firms is too thin relative to the log normal rather than too fat, as had previously been believed.

**Keywords:** Firm size; Zipf plot; Gibrat's Law

**JEL classification:** L11

---

This paper presents new evidence on the size distribution of firms. Like earlier studies, it shows that the log-normal distribution fits the data well except for the upper tail. However, in contrast to earlier studies, we find that there is too little mass in the upper tail, not too much. We demonstrate this point with a statistical technique that has been used rarely in economics, but is more common in physics.<sup>1</sup> The technique, known as a Zipf plot, is a plot of the log of the rank vs. the log of the variable being analyzed.

Let  $(x_1, \dots, x_N)$  be a set of  $N$  observations on a random variable  $x$  for which the cumulative distribution function is  $F(x)$ , and suppose that the observations are ordered from largest to smallest so that the index  $i$  is the rank of  $x_i$ . The Zipf plot of the sample is the graph of  $\ln x_i$  against  $\ln i$ . Because of the ranking,  $i/N = 1 - F(x_i)$ , so

$$\ln i = \ln[1 - F(x_i)] + \ln N. \quad (1)$$

Thus, the log of the rank is simply a transformation of the cumulative distribution function. It accentuates the upper tail of the distribution and therefore makes it easier to detect deviations

\* Corresponding author.

<sup>1</sup> See Gell-Mann (1994, p. 93) for a discussion.

in the upper tail from the theoretical prediction of a particular distribution. Since there has been interest in the upper tail of the size distribution of firms, the Zipf plot is particularly useful for analyzing this question.

The Zipf plot for the log-normal distribution is characterized by

$$\ln i = \ln \left[ 1 - \Phi \left( \frac{\ln x_i - \mu}{\sigma} \right) \right] + \ln N, \quad (2)$$

where  $\mu$  and  $\sigma$  are the mean and standard deviation of  $\ln x_i$ , and  $\Phi$  is the standard normal cumulative distribution function. Solving (2) for  $\ln x_i$  as a function of  $\ln i$  gives

$$\ln x_i = \sigma \Phi^{-1} \left( 1 - \frac{e^{\ln i}}{N} \right) + \mu. \quad (3)$$

The data for this study are the 1993 sales of 4071 manufacturing firms (SIC codes 2000–3999) on Compustat.<sup>2</sup> Fig. 1 shows a histogram of the log of sales with bin sizes equal to  $\sqrt{2}$ . The curve is the normal density function with mean and standard deviation equal to the sample mean and standard deviation of the log of sales. The graph seems to suggest that the

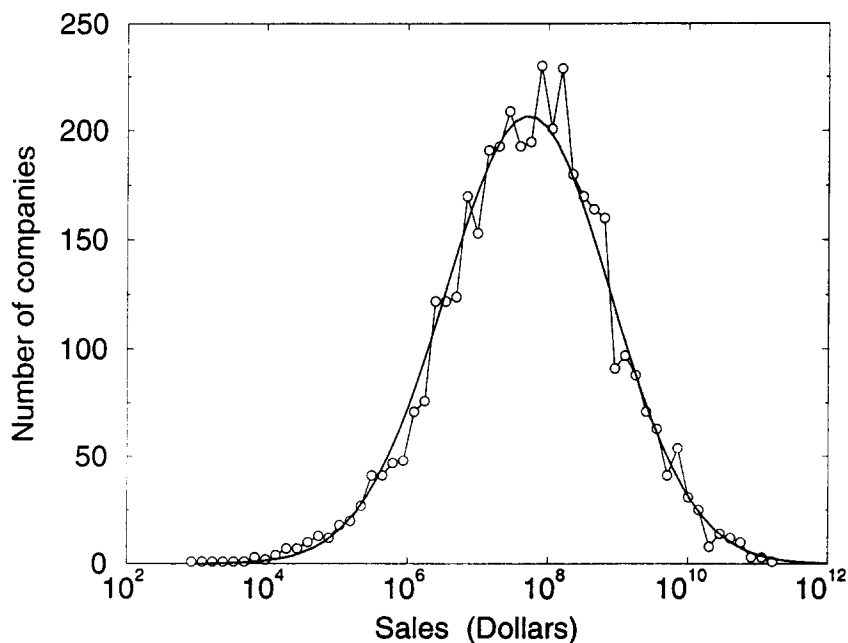


Fig. 1. Distribution of firm size. The circles are a histogram showing the number of firms having 1993 sales of  $X$  dollars as a function of  $\log X$ . The data are for the 4071 Compustat firms in SIC codes 2000–3999. The values of the sales are binned in powers of  $\sqrt{2}$ . The solid curve is a log-normal fit to the data using the mean of the log of sales and the standard deviation of the log of sales as fitting parameters.

<sup>2</sup> Compustat is not, of course, the entire population of firms. In principle, though, it is the entire population of publicly traded firms. While we only report results here for 1993, we have done the analysis for 1975, 1979, 1980, and 1984 and obtained qualitatively similar results.

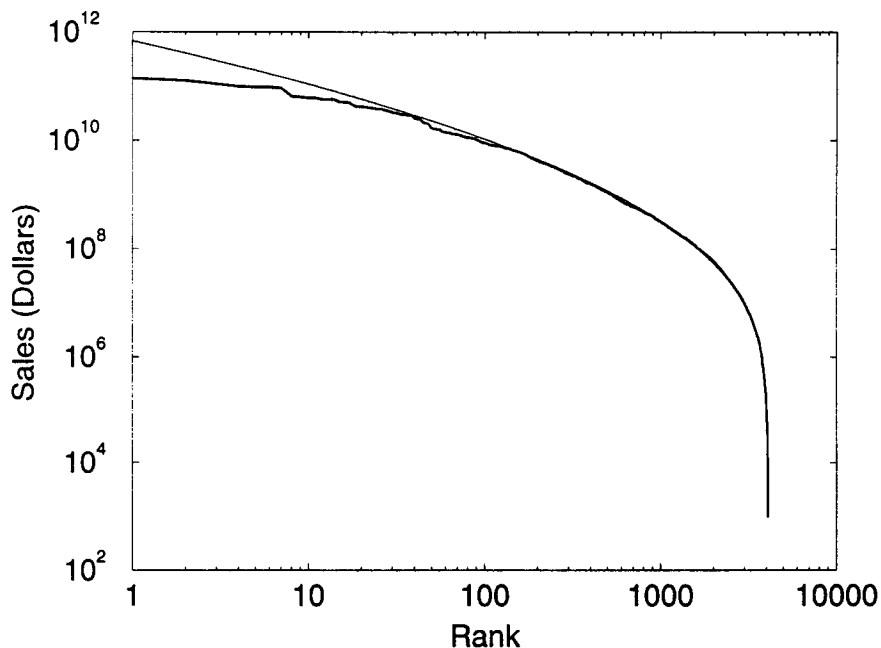


Fig. 2. Zipf plot. The bottom curve is a Zipf plot (double logarithmic plot of sales vs. rank) for the same sample as in Fig. 1. The top curve is a predicted Zipf plot obtained from the log-normal fit shown in Fig. 1.

distribution of the log of sales fits the log normal reasonably well. Fig. 2 shows the Zipf plot along with the theoretical Zipf plot for the log normal. Like the histogram, the Zipf plot suggests that the log normal fits the distribution of sales reasonably well. However, in contrast to the histogram, the Zipf plot makes clear that the sales of the largest firms are smaller than would be the case for a true log normal. The actual Zipf plot lies below the theoretical Zipf plot for roughly the largest 100 firms.

With the aid of the Zipf plot, the deviations from the log normal can also be seen in Fig. 1. First, the three points on the right lie slightly below the best fitting density function. The main source of deviation is, however, that the upper tail should contain additional firms. The largest firm in the sample, General Motors, has sales of \$136 billion. The natural log of GM's sales is 25.63. The mean of the natural log of sales is 17.76 and the standard deviation is 2.72. Thus, the natural log of GM's sales is 2.90 standard deviations above the mean. The probability that an observation from a standard normal distribution exceeds 2.90 is 0.0019. Multiplying this probability by the number of firms (other than GM) in the sample (4070) gives 7.73, which is the expected number of firms with sales greater than \$136 billion. If the distribution were log normal, therefore, we would expect GM's level of sales to be the eighth or ninth largest, not the first. All that would be needed for the size distribution of firms to be log normal would be seven firms larger than GM!

This deviation from log normality is statistically significant. Under the null hypothesis of log normality, the number of firms with sales greater than \$136 billion has a binomial distribution with  $p = 0.0019$  and  $N = 4070$ . The variance is, therefore,  $4070 \times 0.0019 \times 0.9981 = 7.72$ ; and the standard deviation is 2.78. Thus, the actual number of firms with sales greater than

\$136 billion, which is 0, is 2.78 standard deviations less than the expected number, which is 7.73.<sup>3</sup> The probability that none of the 4070 firms other than GM would have sales greater than \$136 billion is  $(1 - 0.0019)^{4070} = 0.00043$ , which is substantially below any conventional standard for significance.<sup>4</sup>

These results are of interest because of their implications for the literature on the dynamics of firm growth. Gibrat (1931) showed that if the distribution of growth rates is independent of firm size, the static distribution of firm size would approach the log normal. In an early empirical test using British data, Hart and Prais (1956) found evidence both that the log normal fits the distribution of firm sizes reasonably well and that the growth rates of firms were independent of initial size. They found, however, statistically significant deviations of the distribution from the log normal by estimating the third and fourth moments of the distribution. The distributions were ‘somewhat skewed to the right and slightly leptokurtotic.’<sup>5</sup> Quandt (1966) proposed four tests of the distribution of firm size. He was able to reject log normality for the Fortune 500 in both 1955 and 1960 with each of the four tests he used. Although he was able to reject every distribution he tested for the Fortune 500 with at least one test for at least one of the years, the two Pareto distributions and the Champernowne generally fit better than the log normal. In summarizing the literature, Hall (1987) wrote: ‘The size distribution of firms conforms fairly well to the log normal, with possibly some skewness to the right’ (p. 584). Thus, the results here may suggest a qualitative change in the size distribution in firms from the earlier time periods used in those studies.<sup>6</sup>

## Acknowledgments

We have benefited from conversations with Glenn Loury, Jeff Miron, and Martha Schary and from a referee’s comments.

## References

- Gell-Mann, M., 1994, *The quark and the jaguar* (W.H. Freeman, New York).
- Gibrat, R., 1931, *Les inégalités économiques* (Sirey, Paris).

<sup>3</sup> Because  $p$  is small,  $N$  is large, and  $Np$  is moderate, the distribution of the number of firms with sales greater than \$136 billion is approximately Poisson. The fact that the standard deviation and the number of standard deviations from the mean are (approximately) equal is due to the well-known result that the variance of a Poisson distribution equals the mean.

<sup>4</sup> Even if there were one firm with sales larger than \$136 billion, log normality could be rejected at the 1% level. Log normality could be nearly rejected at the 5% level if there were three firms with sales above \$136 billion. (The  $p$ -value for three firms is 0.051.)

<sup>5</sup> The skewness of the log of sales in our sample is  $-4.01$ . The kurtosis is 176.7, which is 3.22 times the square of the variance. This ratio is only slightly above the theoretical value of 3 for a normal distribution.

<sup>6</sup> An alternative explanation is that the earlier studies by Simon and Bonini (1958) and Quandt (1966) used the Fortune 500 as their samples.

- Hall, B.H., 1987, The relationship between firm size and firm growth in the U.S. manufacturing sector, *The Journal of Industrial Economics* 35, 583–606.
- Hart, P.E. and S.J. Prais, 1956, The analysis of business concentration: A statistical approach, *Journal of the Royal Statistical Society, Series A*, 119, 150–181.
- 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.