FSD, power-laws and truncation point

Agustí Segarra
Mercedes Teruel

ABSTRACT

This paper analyzes the behaviour of firm size distributions (FSDs) in relation to the truncation point. In the real markets, FSD presents numerous small firms and a few large firms. Many empirical size distributions in economics and other fields exhibit a power law in the upper tail (income, city size, firm size). Multiplicative models à la Gibrat (1931) present FSDs close to lognormal distributions and with upper tails close to Pareto or Zipf distributions. For an exhaustive sample of Spanish manufacturing firms, we show the existence of a non-constant power-law distribution that depends on the sampling size we consider. Furthermore, the FSD of employees is more sensitive to firm age than the FSD of sales.

JEL classification: L11, C16
Key words: Firm growth, FSD, power-laws, truncation sample, Spanish manufacturing firms

Contact: agusti.segarra@urv.cat, mercedes.teruel@urv.cat
Research Group of Industry and Territory
Department of Economics
Rovira i Virgili University
Av. Universitat, 1; 43204 – Reus, Spain
Tel. + 34 977 759 816 Fax + 34 977 300 661

Acknowledgements
We are grateful to the Spanish Ministry of Industry, Tourism and Trade for financial support. Veronica Gombau provided excellent research assistance. The usual disclaimers apply.

1. Introduction

Recently, in the field of Industrial Organization a debate has emerged about whether firm size distributions are best modelled by a power-law distribution or a lognormal distribution. Multiplicative models have become a standard reference in Industrial Organization (Gibrat, 1931). In these models the evolution of an individual firm follows a sequence of stochastic multiplicative shocks. The connection between multiplicative processes and the lognormal distribution was made by Robert Gibrat, who built upon the work by astronomer Jacobus C. Kapteyn (1903) and the mathematician Donald McAlister (1879). Kapteyn developed a multiplicative process that gives rise to an asymptotically lognormal distribution, and McAlister described the lognormal distribution around 1879. In his PhD thesis Gibrat (1931) observed that the firm size distribution (FSD) was close to the lognormal distribution, and he concluded that firm growth rates follow a random process.

Several models of proportional growth were subsequently introduced in economics to explain firm growth rates and market dynamics. In particular, Simon (1955) and Ijiri and Simon (1977) extended Gibrat's model by introducing an entry process in which the number of firms changes and rises over time. Ijiri and Simon (1977) demonstrated that the largest firms are close to the Pareto distribution in the upper tail of the FSD.

The aim of this paper is to improve understanding of how FSDs behave in relation to the power law they are defined by and their relationship with the sample size. In recent years access to extensive databases containing more firms, especially small businesses, has facilitated new analytical perspectives. In the late 1990s, the availability of U.S. Census Department data on the entire universe of U.S. businesses has given rise to new approaches (Axtell, 2001; Teitelbaum and Axtell, 2005; Bottazzi et al, 2006).

These studies analyse FSDs in two stages. First, they analyse the shape of FSDs (in logs) at an aggregated and sectoral level by applying kernel densities. Second, they observe the differences in the FSDs classified by firm size and sector. In this second stage, they apply two measures to calibrate the number of modes and the sectoral concentration. In general, the FSDs approximate more to a Pareto distribution with an exponent near unity. In other words, FSDs approach a Zipf distribution, which is very common in the analysis of the distribution of city sizes (Gabaix, 1999; Eeckhout, 2004).

Some recent examples can be found in Axtell (2001) and Kaizoji et al. (2005). Axtell (2001) used an exhaustive Business Master File for 1997. He observed that FSDs are well-approximated by the Pareto distribution with exponent near unity -the so-called Zipf distribution- throughout the range of firm sizes. Recently, Kaizoji et al. (2005) used the Bloomberg database of multinational firms in 1995 and 2003 to analyze FSDs in terms of sales and total assets of Japanese and US companies. They found that the FSD of US firms is approximately lognormal, in agreement with Gibrat's model. In contrast, the FSD of Japanese firms is clearly not lognormal, and the upper tail follows the Pareto law, according to the Simon model.

In Industrial Organization, Sutton (1997) and Jovanovic (1982) studied the relation between proportionate growth and size distributions that were not lognormal. Gabaix (1999) and Blank and Solomon (2000) offer a solution to the puzzle of city size distribution and show that proportionate growth processes can generate Zipf's Law at the upper tail. However, the constant debate is whether firms are distributed as power laws (e.g. Pareto, Zipf) or as lognormal (Aitchison and Brown, 1954; Champernowne, 1953; Axtell, 2001; Mitzenmacher, 2004).

Our aim is to determine the power-law relationship between firm rank and firm size. We apply a rolling sample methodology to analyse the effect of including small firms in the estimation. In order to do the empirical calibration we use an extensive database for Spanish manufacturing firms that compiles information of balance sheets from Spanish Mercantile Register in 2001 and 2006.

2. Power laws and firm size distribution

Several studies have analysed the relationship between firm rank and firm size. This literature is related to market structure and firm growth (Steindl, 1965; Ijiri and Simon, 1977; Jovanovic, 1982; Sutton, 1997; Amaral et al., 1997).

In general, a non-negative random variable X describes a power-law distribution if the *complementary cumulative distribution function* (ccdf), or $\Pr[X \geq x]$, satisfies

$$Pr[X \geq x] \sim cx^{-\alpha}$$

where constants $c > 0$ and $\alpha > 0$.[1] It is easy to observe that the Pareto distribution is a power law that satisfies,

$$\Pr[X \geq x] = \left(\frac{x}{k}\right)^{-\alpha}$$

for some $\alpha > 0$ and $k > 0$. Note that Pareto distribution requires $X \geq k$. If α falls in the range $0 < \alpha < 2$, then X has infinite variance. If $\alpha \leq 1$, then X also has an infinite mean. When $\alpha = 1$ this distribution is known as Zipf's Law (Zipf, 1949). The Zipf distribution is a special case of the Pareto distribution and presents the usual behaviour of power-law distributions (Richiardi, 2004). However, instead of considering one distribution defining the FSD, recent empirical literature observes a mixture among lognormal and other power laws. Gibrat's model (1931) showed the link between a lognormal FSD and power laws in both tails. Recently, Solomon and Levy (1996)

[1] For details, see Mitzenmacher (2004).

showed that a power law can also be obtained by adding a reflection condition to the Gibrat model (i.e. by assuming that firm size is bounded from below to a threshold proportional to the average firm size).

Along the same lines, Reed (2003) and Mitzenmacher (2004) observed a double Pareto FSD. Reed (2003) provided a distribution that is closer to lognormal for large samples and closer to the Pareto distribution in both tails for large or small values. He calls this a double Pareto distribution. Given the empirical literature is not clear if lognormal or power-law distributions are better models to define the shape of FSD.

In order to analyse the FSD of Spanish manufacturing firms we propose two approaches. First, we calibrate the Zipf distribution and, second, we analyze the adjustment of the full sample to lognormal distribution. We can express the Zipf distribution as,

$$r = N\,(1 - P(S)) = N\,\left(S/k\right)^{\alpha}$$

where N is the number of observations, r is the rank and S is the firm size measured in number of employees. Zipf distribution is usually estimated by ordinary least squares and the regression adopts the following equation,

$$\ln r = K - \alpha \ln S + \varepsilon$$

where K and α are the coefficients to be estimated and ε is a random error. Depending on α, there are three possible results. First, if α is closer to 1 the FSD is a Zipf distribution. Second, if α is larger than unity the relationship between firm size and rank is superlinear. In other words, firm sizes diminish less than the quotient between the largest firm size and the rank that a firm occupies in the distribution. Third, if α is smaller than unity, the relationship between firm size and rank is sublinear. In other words, firm sizes diminish more than the quotient between the largest firm size and the rank[2].

3. Rolling sample results for Spanish manufacturing firms

The data for this study is for Spanish manufacturing firms in 2001 and 2006. The database contains information about the balance sheets in the Spanish Mercantile Register and our sample consists of those firms with more than two employees. In order to analyse the power law of the FSD, we consider the number of employees and sales. The results in Table 1 show a negative relationship between the estimated coefficient and sample size ($d\alpha/dN < 0$). This means that small samples of large firms yield higher coefficients ($\alpha>1$) than large samples that also include smaller firms,

---

[2] In the field of urban systems, Eeckhout (2004) demonstrates that if a variable adopts a lognormal distribution, the value of the parameter α from the Pareto distribution increases when the *truncation* size increases ($d\alpha/dS > 0$) and decreases when the sample size of population increases ($d\alpha/dN < 0$). Similar results are obtained by González-Val (2006) for cities and metropolitan areas in the USA during the 20th century.

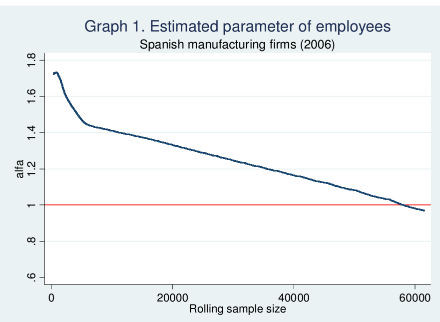
regardless of the variable (employees or sales). For small samples, the relationship is superlinear. But at some point of the estimation, the coefficient is equal to unity; however, when the sample increases, the parameter decreases. As a consequence, the second largest firm is larger than half the size of the first. A plausible explanation is that the largest firms are more homogeneous, while small firms are more heterogenous.[3]

Table 1. Regression results on Zipf's Law

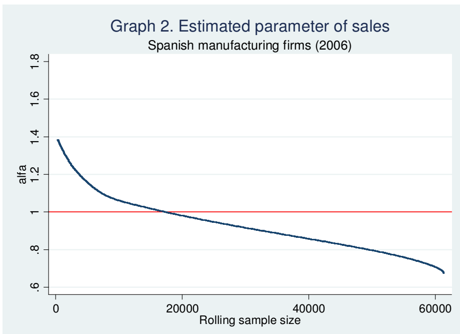
	Employees				Sales			
	2001		2006		2001		2006	
	α	R^2	α	R^2	α	R^2	α	R^2
100	1.5496 (0.0259)*	0.9736	1.5590 (0.0223)*	0.9805	1.3336 (0.0200)*	0.9787	1.2847 (0.0152)*	0.9866
500	1.7201 (0.0072)*	0.9913	1.7310 (0.0064)*	0.9933	1.3589 (0.0040)*	0.9956	1.3677 (0.0041)*	0.9955
1000	1.6756 (0.0038)*	0.9949	1.7262 (0.0032)*	0.9966	1.3355 (0.0021)*	0.9975	1.3335 (0.0023)*	0.9971
5000	1.4414 (0.0017)*	0.9929	1.4690 (0.0020)*	0.9911	1.1706 (0.0014)*	0.9932	1.1567 (0.0014)*	0.9928
10000	1.3795 (0.0010)*	0.9947	1.4101 (0.0011)*	0.9940	1.0777 (0.0010)*	0.9910	1.0608 (0.0010)*	0.9904
20000	1.2892 (0.0008)*	0.9926	1.3314 (0.0008)*	0.9934	0.9725 (0.0008)*	0.9868	0.9806 (0.0007)*	0.9897
40000	1.1001 (0.0009)*	0.9758	1.1638 (0.0008)*	0.9813	0.8190 (0.0007)*	0.9713	0.8577 (0.0006)*	0.9804
Total	0.9449 (0.0009)*	0.9476	0.9697 (0.0009)*	0.9503	0.6608 (0.0009)*	0.9131	0.6749 (0.0008)*	0.9209
N	54490		61455		54382		61322	
Truncation point	49,944		58038		16,997		17311	

* Significant at 1%.

However, the truncation point differs between both variables and regardless of the year. Graphs 1 and 2 show how the coefficient decreases when small firms are added. However, the truncation point differs between employees and sales. For example in 2001 the coefficient equal to 1 is reached in the rank position equal to 49,944 for the employees, while it is reached in the rank position equal to 16,997 for sales.



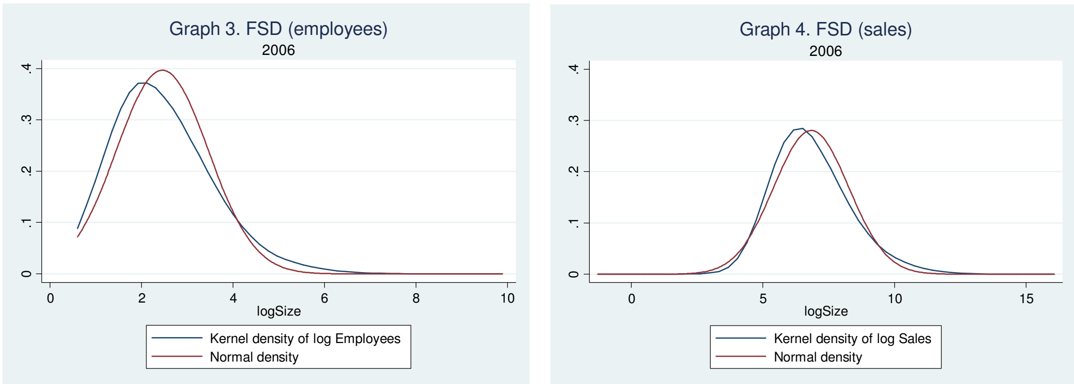
Graph 1. Estimated parameter of employees



Graph 2. Estimated parameter of sales

[3] In the context of city size, Gabaix (1999a, b) explained this result as the consequence of economies of scale. First, the largest cities enjoy similar economies of scale, degree of diversity and productivity, so their city size does not differ very much. Second, the inclusion of small cities decreases coefficient α. In our case, a crucial variable that may affect firm size is age. Table A-1 shows how firm age depends on the rank size. In fact, there is a positive relationship between firm size and age.

The pattern of the parameter of employees and sales may be different because the FSD of both variables behaves differently. Graphs 3 and 4 report the estimated FSD of the log employees and sales in 2006. We have added the normal density distribution so that the deviation can be compared. The two distributions show significant differences. The FSD of employees differs from the normal density while the shape of the log sales is much more similar to normal density, although it is slightly biased towards the right. However, both graphs show that the density in the largest value is higher than the normal density expected in Zipf's Law. In other words, the upper tail concentrates a higher density than in the normal density. Consequently, a group of large firms are performing better that is to be expected.



First of all, the above graphs reveal considerable, widespread heterogeneity across firms, which produces a skewed FSD. However, we observe differences between employees and sales (i.e. between corporate performances and corporate characteristics). Corporate performance seems to approach more to normal density. Second, the graphs confirm an upward bias of samples among the smallest and the largest firms (kernel densities concentrate more probability in the extreme density than in the normal density).

Table 2. The relationship between the log of the estimated Pareto exponent and the log of the sample size (2001)

	γ	δ	R^2
	2001		
Employees	1.4087 (0.0026)*	-0.1224 (0.0003)*	0.7989
Sales	1.4205 (0.0028)*	-0.1515 (0.0003)*	0.8395
	2006		
Employees	1.4276 (.0024)*	-0.1211 (0.0002)*	0.8082
Sales	1.3273 (.0023)*	-.1407 (.0002)*	0.8608
* significant at 1%.			

In order to check the relationship between the sample size and the parameter of rank-size Law, we ran a regression between the estimated exponent $(\hat{\alpha})$ and the sample size (SS). For this analysis we ran the following equation and present the results in Table 2.

$$\log(\hat{\alpha}_i) = \gamma - \delta \log(SS_i) + \varepsilon$$

Table 2 shows that including a larger number of observations using the rolling sample method negatively affects the estimated exponents $(\hat{\alpha})$. The results confirm our expectation. FSD is greatly affected by sample size. As a consequence, the validity of Zipf's law depends on the sample size used in a study and increasing sample size has a negative impact on the parameter of the power law.

Zipf's Law and firm age

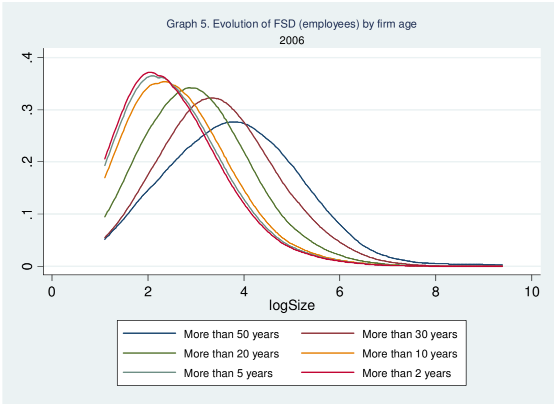
As we have pointed out above, the differences in α between large and small firms may be because small firms are more heterogeneous than large firms. For this reason, we may wonder how a variable such as firm age affects the evolution of the parameter.

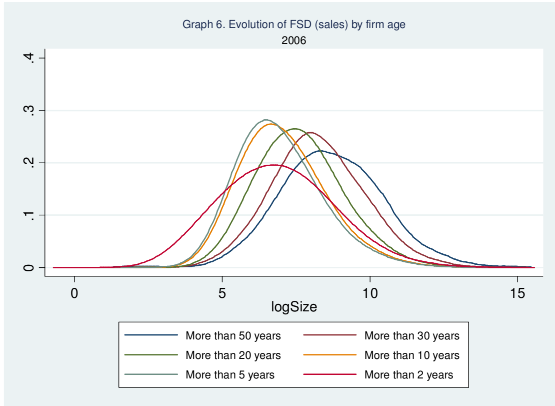
Table 3. Regression results on Zipf's Law using number of employees and sales (2001 and 2006)

	Employees				Sales			
	2001		2006		2001		2006	
More than…	α	R^2	α	R^2	α	R^2	α	R^2
50 years	1.2152 (0.0082)*	0.9758	1.2450 (0.0081)*	0.9748	0.8990 (0.0098)*	0.9396	0.8913 (0.0092)*	0.9385
30 years	1.1771 (0.0036)*	0.9739	1.1881 (0.0031)*	0.9732	0.8788 (0.0035)*	0.9556	0.8577 (0.0030)*	0.9534
20 years	1.1214 (0.0022)*	0.9692	1.0952 (0.0018)*	0.9660	0.8295 (0.0020)*	0.9551	0.7903 (0.0015)*	0.9533
10 years	1.0232 (0.0014)*	0.9581	1.0092 (0.0011)*	0.9558	0.7467 (0.0012)*	0.9412	0.7150 (0.0010)*	0.9359
5 years	0.9725 (0.0011)*	0.9516	.09884 (0.0010)*	0.9526	0.6969 (0.0010)*	0.9281	0.6954 (0.0009)*	0.9295
2 years	0.9547 (0.0010)*	0.9494	0.9778 (0.0009)*	0.9514	0.6789 (0.0009)*	0.9228	0.6853 (0.0008)*	0.9257

* Significant at 1%.

As we expected, there is a positive relationship between the parameter α and firm age. This confirms the fact that young firms are smaller than Zipf's Law would suggest. However, there are differences between the two variables. First, the variable of employees has a superlinear relationship that changes into a sublinear relationship. Second, firm sales continuously show a sublinear relationship regardless of firm age. It seems that firm age affects the number of employees but not so significantly the distribution of sales.





Graphs 5 and 6 show the evolution of FSD with firm age. They reveal two characteristics. First, FSD evolves towards the right when the oldest firms are considered. Second, firm age affects the evolution of the FSD much more in terms of employees than in sales. These results show that there is a rich statistical structure in firm dynamics. Indeed, there will probably be a correlation mechanism between FSD and firm age. This mechanism stabilizes FSD over time.

4. Conclusions

Using the balance sheets of Spanish manufacturing firms, we aim to analyse the power law defining the FSD of employees and sales. The significant high values obtained when estimating parameter α indicates that there is no constant '*power law*' between firm size and firm rank, regardless of the variable. In fact, there is a negative relationship between α and the number of observations in the estimation. Our results are confirmed by the kernel density of both variables: their upper tail has a greater density than in the normal FSD, which causes a super linear relationship.

We conclude with several statements. First, the diversity of results obtained in the literature may reflect differences in sample size. We believe that this conclusion is in line with much of the theoretical and empirical literature available on the topic, which points out that including small firms in the FSD tends to be inversely related to the parameter of the power law. Second, the different patterns of the largest and the smallest firms may be the result of the firms' characteristics (economies of scale, firm age and productivity levels): the largest firms are more homogeneous than the smallest ones. Finally, firm age affects the power-law parameter. The relationship between firm age and the estimated parameter is positive. However, there are significant differences in the estimation of the FSD of employees and sales.

References

Aitchinson, J. and Brown, J. A. C. (1954): "A Synthesis of Engel Curve Theory", *Review of Economic Studies,* 22(1):35-46.

Amaral, L.A. N., Buldyrev, S.V., Havlin, S., Leschhorn, H., Maass, P., Salinger, M.A., Stanley, H.E. and Stanley, M. H. R. (1997): "Scaling behavior in economics: I. Empirical results for company growth", *Journal of Physics*, 7: 621–633.

Axtell, R. L. (2001): "Zipf distribution of US firm sizes", *Science*, 293(5536):1818-1820.

Blank, A. and Salomon, S. (2000): "Power-laws in cities population, financial markets and internet sities (scaling in systems with a variable number of components)", *Physica A*, 287, 279-288.

Bottazzi, G., Secchi, A. and Tamagni, F. (2006): "Financial Fragility and Growth Dynamics of Italian Business Firms", Working Paper; Sant'Anna School of Advanced Studies.

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

Eeckhout (2004): "Gibrat's Law for (All) Cities", *American Economic Review*, 94(5): 1429-1451.

Eeckhout, J. (2004): "Gibrat's Law for (All) Cities", *American Economic Review*, 94:1429-1451.

Gabaix, X. (1999): "Zipf's Law and the Growth of Cities", *The American Economic Review*, 89 (2): 129-132.

Gabaix, X. (1999): "Zipf's Law for Cities: An Explanation", *Quarterly Journal of Economics*, 114: 739-767.

Gibrat, R. (1931): *Les inegalites economiques*. Paris: Libraire du Recueil Sirey.

Gibrat, R. "Une loi des réparations économiques: l'effet proportionnel." Bull. Statist. g´en Fr. 19 (1930), 469.

González-Val, R. (2006): "Evolución de la estructura urbana USA, 1900-2000", Departamento de Análisis Económico, Universidad de Zaragoza, mimeo.

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

Jovanovic, B. (1982): "Selection and Evolution of industry", *Econometrica*, 50, 649-670.

Kaizoji, T.; H. Iyetomi and Y. Ikeda (2005): "Re-examination of the size distribution of firms", *Evolutionary and Institutional Economics Review*, 2: 183-198.

Kapteyn. J.C.(1903): *Skew Frequency Curves in Biology and Statistics*, Astronomical Laboratory, Groningen: Noordhoff.

Solomon, S. and Levy, M. (1996): "Spontaneous Scaling Emergence in Generic Stochastic Systems", *International Journal of Modern Physics C*, 7 (5): 745.

McAlister, D. (1879): "The Law of the Geometric Mean", *Proceedings of the Royal Society*, 29, 367-376.

Mitzenmacher M. (2004): "A Brief History of Generative Models for Power Law and Lognormal Distributions", *Internet Mathematics*, 1(2): 226—251.

Reed, W. J. (2003): "The Pareto law of incomes - an explanation and an extension", *Physica A: Statistical Mechanics and its Applications*, 319: 469-486.

Richiardi, M. (2004a): "Generalizing Gibrat: Reasonable Multiplicative Models of Firm Dynamics", *Journal of Artificial Societies and Social Simulation*, 7(1).

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

Steindl, J. (1965): *Random Processes and the Growth of Firms*, London: Griffin.

Subbotin, M.T. (1923): "On the law of frequency of errors", *Matematicheskii Sbornik*, 31: 296-301.

Sutton, J. (1997): "Gibrat's Legacy", *Journal of Economic Literature*, 35, 40-59.

Teitelbaum, D. and Axtell, R. (2005): "Firm Size Dynamics of Industries: Stochastic Growth Processes, Large Fluctuations and the Population of Firms as a Complex System", Small Business Advocacy, United States Government.

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