

HEAVY-TAILS AND THE CENTRAL LIMIT THEOREM IN BUSINESS SURVEYS

Jack Lothian ¹

ABSTRACT

Heavy-tailed distributions are ubiquitous in business surveys. While the functional form of these distributions varies, most of the 'best-fit' are positive-definite right-skewed sub-exponential distributions with finite first moments. This paper explores the empirical size distributions of Canadian corporations over a 25-year period. A lognormal distribution was an excellent fit for the size distribution, a log-Laplace for the growth rates, and a log-Subbotin for the ratio of 2 size measures. The Central Limit Theorem applies to the size measure but possibly not to the growth rates and it probably does not hold for some ratios of economic variables.

KEY WORDS: Central Limit Theorem, firm size distribution, log-normal distribution.

RÉSUMÉ

Les distributions à queues lourdes sont omniprésentes dans les enquêtes auprès des entreprises. Alors que la forme fonctionnelle de ces distributions varie, la plupart des distributions « les mieux ajustées » pour les enquêtes auprès des entreprises sont des distributions sous-exponentielles définies positives, asymétriques sur la droite et ayant des premiers moments finis. Cet article explore les distributions empiriques des tailles des corporations canadiennes sur une période de 25 ans. Une distribution log-normale s'est avérée la distribution « la mieux ajustée » pour la distribution de la taille, alors qu'une distribution log-Laplace fut un bon choix pour les taux de croissance, et que pour le ratio de deux mesures de taille, la distribution log-Subbotin fut le meilleur choix. On notera que le théorème central limite s'applique aux mesures de taille, mais possiblement pas aux taux de croissances, et que le théorème ne semble pas être approprié pour certains ratios de variables économiques.

MOTS CLÉS : Distribution log-normale; distribution par taille d'entreprises; théorème central limite.

1. INTRODUCTION

Heavy-tailed right-skewed distributions are ubiquitous in the social and hard sciences (Reed and Hughes 2002). They generally appear in systems of complex interacting units and they are often seen in conjunction with collective self-organizing behavior. Contrary to popular opinion, heavy-tailed distributions are more widespread than normal distributions (Limpert, Stahl and Abbt 2001, Willinger, Alderson, Doyle and Lun 2004).

The apparent functional form of these heavy-tailed right-skewed distributions is quite varied and has generated much debate. Depending on the data set, researchers have found 'best-fits' to be the Pareto, double-Pareto, log-normal, Weibull, Gamma, Beta, log-Laplace, t , F , Chi, Subbotin, log-Subbotin, Lévy-power, mixed normal, mixed log-normal and mixed exponential distributions. The most common choices are the log-normal and Pareto distributions but these choices represent a plurality. Despite the variety of shapes fitted, there are commonalities among the choices. The overwhelming majority of the 'best-fit' distributions are positive-definite right-skewed sub-exponential distributions with a finite first moment. Sub-exponential distributions (Goldie and Kluppelberg 1998) are heavy-tailed distributions whose tails decrease more slowly than any exponential tail. (i.e. $(1 - F(x))/e^{-\varepsilon x} \rightarrow \infty, \forall \varepsilon > 0$)

¹ Jack Lothian, Business Survey Methods Division, Statistics Canada, 17-J R.H. Coats Building, 100 Tunney's Pasture Driveway, Ottawa, K1A 0T6, Lothian@statcan.ca

The log-normal distribution is a good first-order approximation of the observed distributions and it has desirable characteristics. It is a right-skewed positive-definite sub-exponential distribution with finite moments that have tractable functional forms. The Central Limit Theorem (CLT) applies to the log-normal distribution. (Ijiri and Simon 1964)

In economic literature (Amemiya and Boskin 1974, Bartels 1977), this fact has long been recognized and the first-order assumption of log-normality has been widely accepted for about a century. In addition, econometricians generally assume that economic variables are multiplicatively related. This assumption arises from basic theory because cascade and proportional effects are assumed to dominate economic relations. In the hard sciences, these type of cascade systems often produce proportional effects which in turn create multiplicatively related variables and self-organizing behavior. The physical, biological and economic sciences have all found evidence that size distributions are generated by a class of stochastic processes related to the process² $dY = Y(\mu dt + \sigma \varepsilon^t)$, i.e., a geometric Brownian motion process (Sutton 1997). The limiting distribution for this process is known to be a log-normal distribution (Reed 2001). In economics, it is assumed this class of processes generates wealth distributions, size distributions of returns in financial markets and firm size distributions. This relationship implies log-normality and a multiplicative stochastic process of the form $Y_t = \alpha Y_{t-1} (\varepsilon_t)^\phi$.

Economic literature is dominated by assumptions of multiplicative relations and distributions that are log-normal. In contrast, statistical literature is dominated by assumptions of additivity, normality, linearity and the CLT. The common practices of statisticians are in stark contrast to other scientific fields where heavy tails are encountered. Physical and social scientists see the observed heavy-tailed distribution as a fundamental property that must be confronted and explained by all predictive models whereas statisticians believe that the CLT makes this phenomena a trivial side effect that can be ignored.

The pervasive use of the CLT in business surveys arises from a number of facts. First, the principal objective of most business surveys is measuring the universe summation of some variable such as output which naturally leads to the use of the mean as an estimator, which in turn implies an acceptance of the CLT. Second, within each business unit the observed size characteristics are generated through sub-aggregations. Thus, the observed values should have a distribution equivalent to the distribution of the sample mean. Logically, the CLT should be exerting its powerful influence. Thirdly, most practitioners are aware that to a first order approximation the distribution of business units is log-normal and the CLT applies to the log-normal distribution. Lastly, an acceptance of the CLT allows one to efficiently estimate the required summation using classical sampling techniques. Without the CLT, business survey methodologist would be hard-put to find feasible solutions.

This paper reviews 25 years of an annual censuses of incorporated Canadian businesses in order to explore the veracity of the paradigm (linearity, additivity, normality and the CLT) underlying current business survey practices.

2. THE EMPIRICAL FINDINGS

2.1 The Distribution of Corporate Revenues 1975-99

When one examines the distribution of corporate revenues in any reference year, one sees that it is extremely long-tailed. Figure 1 shows the distribution of corporate revenue sizes for the year 1988. It is clearly non-normal and skewed. Following our assumption that to a first order approximation the distribution is log-normal, let us look at the distribution of the log transformation (Figure 2). The distribution is symmetric and the distribution appears to be remarkably similar to a normal distribution.

² In economics this is referred to as the law of proportional effect or Gibrat's law. (Gibrat 1931)

Figure 1

Distribution (Histogram) of Firms by Revenue Size, 1988

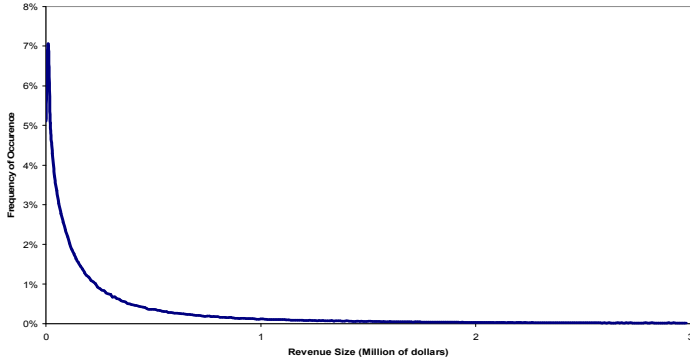


Figure 2

Distribution of Log of Firm Revenues Adjusted for Rounding Effect, 1988

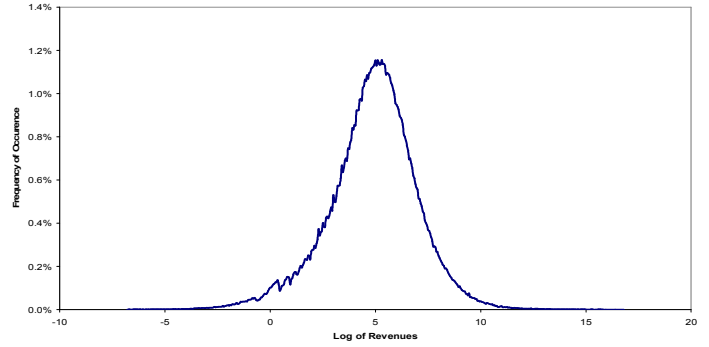
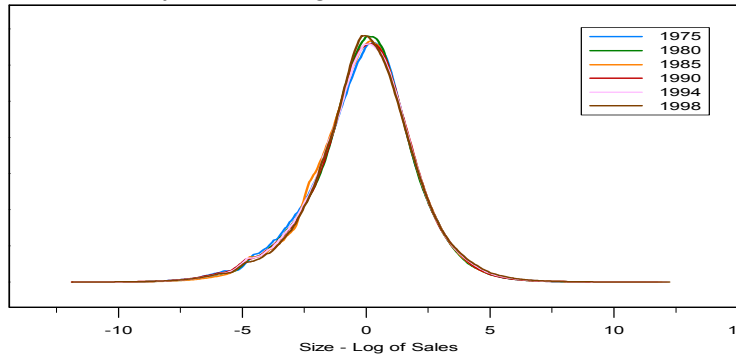


Figure 3 shows the evolution in the shape of the distribution over time³.

Figure 3

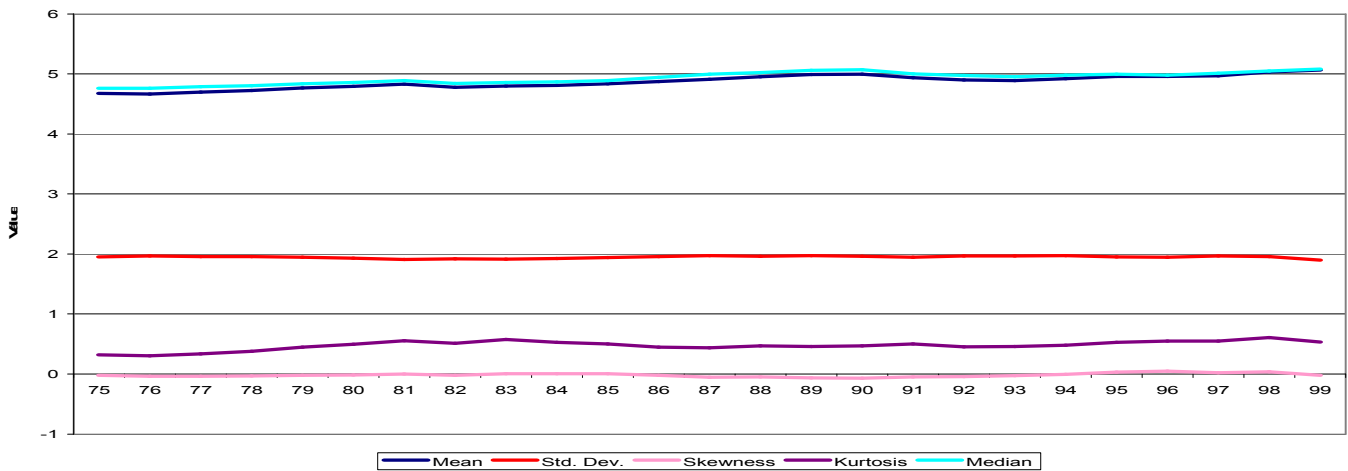
Distribution of Log of Sales - All incorporated firms - Selected Years
Adjusted for Rounding - Kernel smoothed - Mean removed



The stability of the distribution over the 25-year period is remarkable. The distribution for the total universe of firms appears to be a stable log-normal distribution over the full 25 years. Figure 4 shows the parameters estimates for this distribution over the 25 years.

Figure 4

Parameter Estimates - Log of Sales Distribution



³ All years were adjusted for rounding, binned using a bin size of 0.05, then the mean was subtracted and the distribution was smoothed using the default kernel smoothing method in S-PLUS.

Figure 4 shows that the standard deviation (σ) is flat with an approximate value of 2, plus the skewness is zero for all years. The stability plus the almost perfect normal distributional shape is unexpected and extraordinary. Clearly log-normality is a stylized fact for the distribution of Canadian corporate revenues. The large σ value of 2 is troubling because it implies that moderately large sample sizes are required for invoking the CLT.

It should be noted that the empirical results are consistent with an assumption of a Geometric Brownian Motion (GBM) process driving corporate growth. A GBM process will generate the observed empirical distribution, but there are other possible generating processes. Finally, the fact that a log-transformation creates a stable-symmetric-normal-type distribution strongly suggests that the multiplicative rather than the additive form of the CLT may be at work here.

These findings are consistent with a large body of empirical research on the size distribution of firms. Recently, the study of Amaral (Amaral, et al. 1997) of US manufacturing firms found similar results over a 20-year period. More recently, Axtell (Axtell 2001) found the same result with a larger set of US data.

2.2 The Distribution of the growth rates in Corporate Revenues 1976-98

Next, let us look at the distribution of the growth rates in corporate revenues at the corporate level. Figure 5 shows the evolving shape of this distribution over the 25-year period. As in the case of revenues, the untransformed distribution is highly skewed to the right, but the log-transform shows a symmetric distribution that is an extremely close fit to a log-Laplace distribution. Again, the stability of the distribution over the 25-year period is remarkable and again, it is the mean that is changing through time while the shape parameter ϕ remains fixed. Figure 6 shows the estimates of parameters of the Laplace distribution through time. The distribution of the growth rates (Y_t / Y_{t-1}) appears to be a stable log-Laplace distribution. Given the fact that the size distribution is log-normal with a fixed standard deviation σ , it is no surprise that the growth rate is a stable Laplace distribution with a fixed shape parameter⁴.

Figure 5

Distribution of Log of Growth in Sales - Selected Years
Rounding Adjustments - Kernel Smoothed - Mean Removed

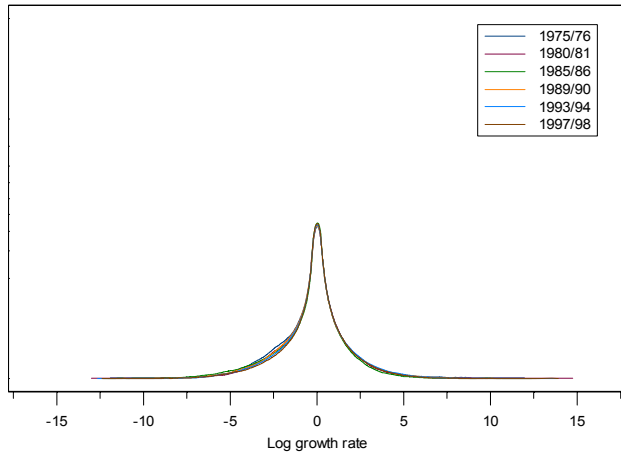
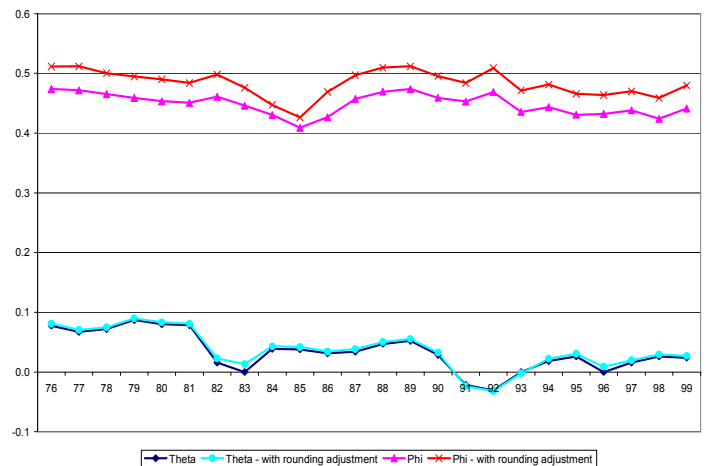


Figure 6

Parameter Estimates of Laplace Distribution



Based upon Figures 5 and 6, a good stochastic model for the revenues of corporation i in year t would be $Y_{i,t} = Y_{i,t-1} \varepsilon_t$ where ε follows a log-Laplace distribution $\varepsilon_t(\theta_t, \phi)$. Under this scenario, the value of ϕ is troubling because if $\phi \geq 0.5$, the variance of the growth rates is undefined. As can be seen from Figure 6, ϕ is on the cusp of the region where the variance becomes infinite and the CLT no longer holds. Note that Y appears to be generated by a GBM process and the results appear to confirm Gibrat's Law (Mansfield 1962).

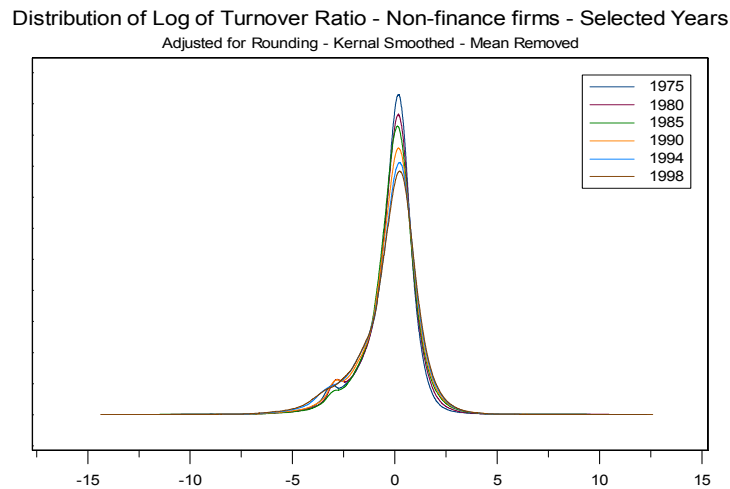
These findings on size growth rates are consistent with many recent empirical studies on firm growth rates. (Amaral, et al. 1997, Okuyama, Takayasu and Takayasu 1999, Axtell 2001, Bottazzi and Secchi 2002)

⁴ Barndorff-Nielsen (Barndorff-Nielsen 1977) suggests a theoretical connection between the log-Laplace distribution and a mixture of log-normal distributions.

2.3 The Distribution of Ratios of Economic Variables 1975-1999

The Capital Asset Turnover ratio is an interesting statistic because it is the ratio of two widely-used measures of size, revenues and assets. Figure 7 shows the evolution of the log of the Turnover ratio over time. While the untransformed ratio is clearly non-normal, the log-transform is a symmetric stable log-Subbotin (Johnson and Kotz 1970) distribution.

Figure 7



Unfortunately, the fitted parameters for the log-Subbotin distribution suggest that the variance of the distributions is infinite. Contrary to a widely held belief, the distributions of financial ratio are ill behaved and non-normal. Despite the fact that both Revenues and Assets are stable log-normal distribution, the distribution of the ratio is a stable log-Subbotin with an infinite variance. This suggests that estimators of the ratio Revenues over Assets at year t that are based upon the CLT are not valid, and the variance estimates in particular are not reliable.

3. COMMENTS

The empirical evidence shows that the distributions underlying business surveys are sub-exponential with a finite first moment, but in some cases the second moment does not exist. In addition, multiplicative relationships appear to dominate economic data and the multiplicative CLT appears to be the most appropriate form of the CLT⁵. Interestingly, in the log-transform, all moments of the distribution are finite and the observed distributions are well behaved and tractable.

Despite these findings, the author believes that under most circumstances, classical statistical theory can still be applied to business surveys. A good example is the use of size stratification in surveys of business units. Defining a take-all stratum that includes all the large firms in the right-hand heavy tail significantly normalizes the remaining units in the survey. This allows one to use classical sampling strategies for medium and small size businesses. Knowing the exact form of the distributions that we are dealing with can help the methodologist develop an efficient and reliable estimation strategy.

REFERENCES

Amaral, L. A. N., et al. (1997), "Scaling Behavior in Economics: I. Empirical Results for Firm Growth," *Journal de Physique I*, **7**, 621-633.

Amemiya, T., and Boskin, M. (1974), "Regression Analysis When the Dependent Variable Is Truncated Lognormal, with an Application to the Determinants of the Duration of Welfare Dependency," *International Economic Review*, **15**, 485-496.

Axtell, R. L. (2001), "Zipf Distribution of U.S. Firm Sizes," *Science*, **293**, 1818-1820.

⁵ The classical Gaussian CLT is not the only version of CLT. There several alternative forms of the theorem including a multiplicative version and Lévy's Generalized Central Limit Theorem (Bartels 1977)

- Barndorff-Nielsen, O. E. (1977), "Exponentially Decreasing Distributions for the Logarithm of Particle Size," *Proceedings of the Royal Society, London (A)*, **353**, 401-419.
- Bartels, R. (1977), "On the Use of Limit Theorem Arguments in Economic Statistics," *American Statistician*, **31**, 85-87.
- Bottazzi, G., and Secchi, A. (2002), "On the Laplace Distribution of Firm Growth Rates," Technical Report 2002/20, Laboratory of Economics and Management (LEM), Sant'Anna School of Advanced Studies.
- Gibrat, R. (1931), *Les Inégalités Économiques; Applications; Aux Inégalités Des Richesses, À La Concentration Des Enterprise, Aux Populations Des Villes, Aux Statistiques Des Families, Etc., D'une Loi Nouvelle, La Loi De L'effet Proportionnel*, Paris: Libraire du Recueil Sirey.
- Goldie, C. M., and Kluppelberg, C. (1998), "Subexponential Distributions," in *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*, eds. R. J. Adler, R. E. Feldman and M. S. Taqqu, Boston: Birkhäuser, pp. 435-459.
- Ijiri, Y., and Simon, H. A. (1964), "Business Firm Growth and Size," *The American Economic Review*, **54**, 77-89.
- Johnson, N. L., and Kotz, S. (1970), *Distributions in Statistics: Continuous Univariate Distributions - 2* (Vol. 2), ed. H. Chernoff, Boston: Houghton Mifflin Company.
- Limpert, E., Stahl, W. A., and Abbt, M. (2001), "Log-Normal Distributions across the Sciences: Keys and Clues," *Bioscience*, **51**, 341-352.
- Mansfield, E. (1962), "Entry, Gibrat's Law, Innovation, and the Growth of Firms," *The American Economic Review*, **52**, 1023-1051.
- Okuyama, K., Takayasu, M., and Takayasu, H. (1999), "Zipf's Law in Income Distribution of Companies," *Physica A: Statistical Mechanics and its Applications*, **269**, 125-131.
- Reed, W. J. (2001), "The Pareto, Zipf and Other Power Laws," *Economics Letters*, **74**, 15-19.
- Reed, W. J., and Hughes, B. D. (2002), "From Gene Families and Genera to Incomes and Internet File Sizes: Why Power-Laws Are So Common in Nature," *Physical Review E*, **66**, 1-4.
- Sutton, J. (1997), "Gibrat's Legacy," *Journal of Economic Literature*, **35**, 40-59.
- Willinger, W., Alderson, D., Doyle, J. C., and Lun, L. (2004), "More "Normal" Than Normal: Scaling Distributions and Complex Systems," in *2004 Winter Simulation Conference*, AQM, pp. 1-12.