The Difference-of-Log-Normals Distribution is Fundamental in Nature

Robert Parham*

Friday 18th November, 2022

Abstract

I describe a hitherto obscure statistical distribution — the Difference-of-Log-Normals — and make the case that it is a fundamental distribution in nature. I show how it arises from first principles, and why it describes a plethora of natural phenomena, such as population growth in standard birth-death models; firm income, growth, and equity returns; city growth; etc. I present implications for the functional form of neo-classical production functions.

JEL classifications: C46, D20, E23, G30, R10

Keywords: Distributions, log-Normal, growth, firm size, city size.

*University of Virginia (robertp@virginia.edu). See https://kn.owled.ge/DLN/ for companion papers.

The Difference-of-Log-Normals distribution, henceforth DLN,[1] is the distribution arising when one subtracts a log-Normal random variable (RV) from another. To define the DLN, consider an RV W such that

$$W = Y_p - Y_n = \exp(X_p) - \exp(X_n) \quad \text{with} \quad \boldsymbol{X} = (X_p, X_n)^T \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{1}$$

in which $\boldsymbol{X}$ is a bi-variate Normal with

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_p \\ \mu_n \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_p^2 & \sigma_p \cdot \sigma_n \cdot \rho_{pn} \\ \sigma_p \cdot \sigma_n \cdot \rho_{pn} & \sigma_n^2 \end{bmatrix} \tag{2}$$

We say W follows the five-parameter DLN distribution, $W \sim \text{DLN}(\mu_p, \sigma_p, \mu_n, \sigma_n, \rho_{pn})$. Parham (2022a) formally characterizes the distribution, deriving its PDF, CDF, closed-form terms for its moments (which are all finite), and a maximum-likelihood estimation procedure for its parameters given data. It also verifies the estimators above and establishes their small-sample properties using extensive Monte-Carlo simulations.

The *sum* of log-Normal RVs has been used in several disciplines including telecommunication, actuary, insurance, and derivative valuation. The DLN, in contrast, is almost completely unexplored. At the time of writing, I was unable to find instances of using it anywhere in the sciences, and only two statistical works considering it, Lo (2012) and Gulisashvili and Tankov (2016). Both papers concentrate on the sum of log-Normals but show their results hold for the difference of log-Normals as well, under some conditions. Nevertheless, this paper and its companions show that the DLN is a *fundamental distribution in nature*, describing a plethora of natural and economic phenomena.

[1] I lack the Chutzpah to name it the Parham distribution, but you can.

To see this, consider first the central limit theorems (CLTs), which state that

$$Y^+ = \lim_{K\to\infty} \frac{1}{K}\Sigma_{i=1}^{K} X_i^+ \sim \mathcal{N} \tag{3}$$

for $X_i^+ \sim \Omega_i^+$ under mild regularity conditions on the Ω_i^+ depending on the version of the CLT used. Put differently, the CLTs state that a phenomenon in nature which is an additive combination of many latent random forces will tend to distribute Normally.

Consider next the multiplicative CLT, sometimes known as "Gibrat's law", which states that

$$Y^* = \lim_{K\to\infty} \left(\Pi_{i=1}^{K} X_i^*\right)^{\frac{1}{K}} \sim \text{log-}\mathcal{N} \tag{4}$$

for $X_i^* > 0 \sim \Omega_i^*$ under similarly mild regularity conditions. Put differently, the multiplicative CLT states that a phenomenon in nature which is a *product* of many latent random forces will tend to distribute log-Normally. Many physical and economic non-negative quantities, such as mass, population count, interest rates, firm sales, and firm value are products of latent random factors and are approximately log-Normally distributed.

Finally, consider a natural phenomenon impacted by two main forces operating in opposite directions, i.e., $W = Y_p - Y_n$. If the two main forces are additive combinations of latent random forces,

$$W^+ = Y_p^+ - Y_n^+ = \lim_{K_p\to\infty} \frac{1}{K_p}\Sigma_{i=1}^{K_p} X_i^+ - \lim_{K_n\to\infty} \frac{1}{K_n}\Sigma_{j=1}^{K_n} X_j^+ \sim \mathcal{N} \tag{5}$$

then clearly the natural phenomenon distributes Normally as well, because the difference of two Normal RV is itself Normal, under mild conditions. But the same is not true if the two main forces are multiplicative combinations of latent random factors. In this case,

$$W^* = Y_p^* - Y_n^* = \lim_{K_p\to\infty} \left(\Pi_{i=1}^{K_p} X_i^*\right)^{\frac{1}{K_p}} - \lim_{K_n\to\infty} \left(\Pi_{j=1}^{K_n} X_j^*\right)^{\frac{1}{K_n}} \sim \text{DLN} \tag{6}$$

because the difference between two log-Normal RVs does not collapse to a log-Normal RV.

This straightforward application of the CLTs implies we should observe the DLN in many natural settings. Prior to presenting it in the data, it is instructive to consider the simple population dynamics ("birth-death") model to gain intuition as to the DLN's emergence.[2]

Denote $N(t)$ the size of the population in some closed natural habitat (with no immigration or emigration) at time t. The population dynamics of the system are described by the ordinary differential equation:

$$\frac{\mathrm{d}N\left(t\right)}{\mathrm{d}t} = b\left(t\right) \cdot N\left(t\right) - d\left(t\right) \cdot N\left(t\right) \tag{7}$$

in which $b\left(t\right)$ and $d\left(t\right)$ are the instantaneous birth and death rates. In discrete time, we can write

$$N_{t+1} = N_t + b_t \cdot N_t - d_t \cdot N_t = N_t\left(1 + b_t - d_t\right) \tag{8}$$

and interpret b_t and d_t as per-period rates. Let $N\left(0\right) = N_0$ be the initial condition of the system, and note that a second boundary condition is $N\left(t\right) = N_t \geq 0$.

Assume b_t and d_t are stochastic, depending on some underlying latent forces such as food availability, climate, predation, etc. We write $\left[b_t, d_t\right] \sim \Phi$, with their joint distribution Φ assumed i.i.d between periods for simplicity. Note that both rates are supported on $(0, \infty)$, as negative birth or death rates are inadmissible. This precludes Φ from being bi-variate Normal. The simplest hypothesis (in the maximum entropy sense) for their distribution is hence the bi-variate log-Normal.[3] We can then write the per-period growth of the population as

$$g_N = \log\left(N_{t+1}\right) - \log\left(N_t\right) = \log\left(1 + b_t - d_t\right) \approx \left(b_t - d_t\right) \sim \text{DLN} \tag{9}$$

with the approximation becoming equality in the continuous case as the length of a discrete period approaches 0. We conclude that population growth in this simple population dynamics model distributes DLN.

[2]The birth-death workhorse model dates as far back as Malthus (1798).

[3]Other always-positive rates (e.g., the federal funds rate) exhibit excellent fit to a log-Normal as well.

To fix ideas, Figure 1 presents several instances of the DLN distribution. Panel (a) presents and contrasts the standard Normal, standard DLN, and standard log-Normal. The standard DLN is defined as DLN(0,1,0,1,0), i.e. the difference between two exponentiated uncorrelated standard Normal RVs. Panel (b) shows the role of the correlation coefficient ρ_{pn}, controlling tail-weight vs. peakedness. Panel (c) repeats the analysis of Panel (b) for a different parametrization common in practical applications, exhibiting the problem of dealing with the DLN's characteristic heavy tails in both the positive and negative directions. Panel (d) presents the data of panel (c) after taking an Inverse Hyperbolic Sine (asinh) transform of the data. The asinh acts as a log transform in both the positive and negative directions, allowing us to observe the characteristic "double Normal" shape of the transformed DLN.

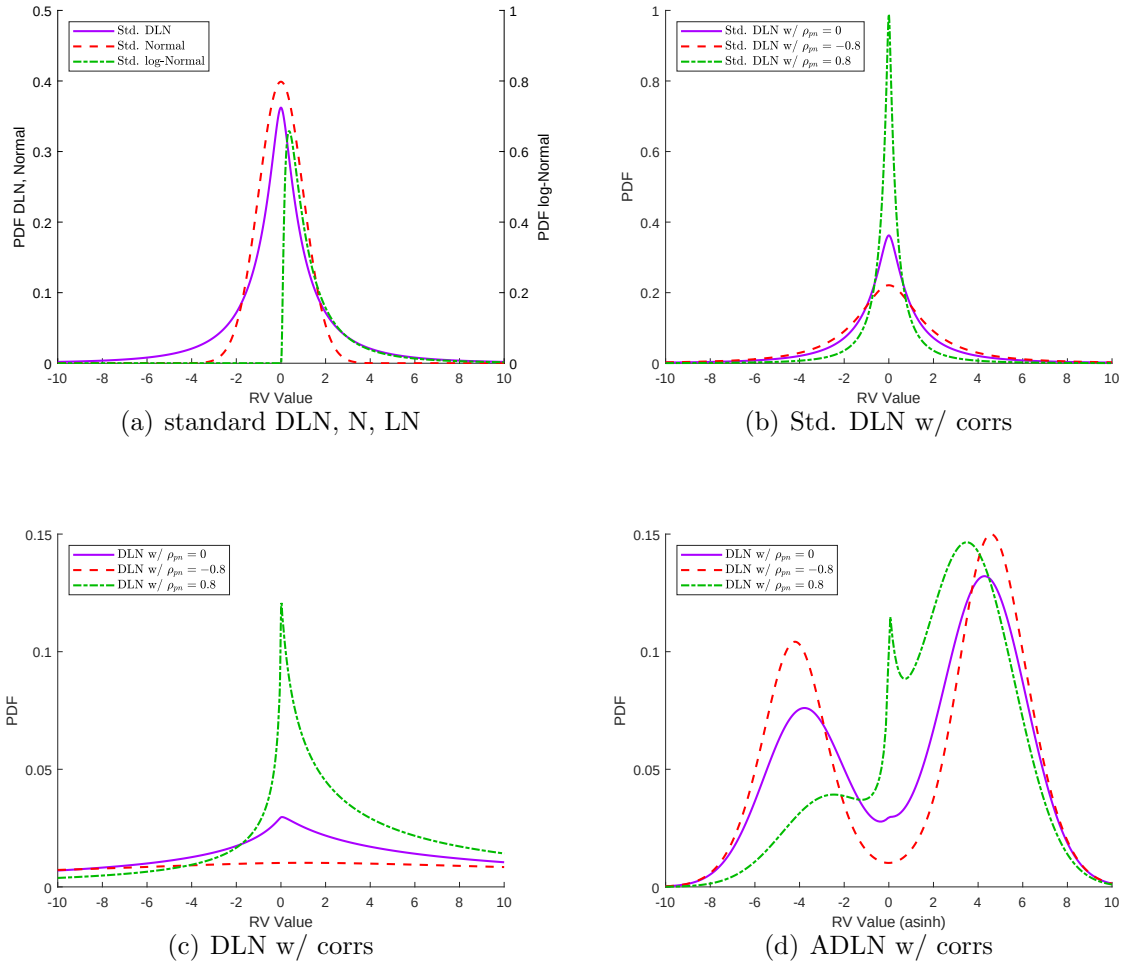
Moving on to the realm of finance, Parham (2022b) empirically analyzes the data on all public US firms in the 50-year period 1970-2019. In a battery of tests, the DLN is shown to be the core distribution driving firm dynamics. The DLN is not rejected for

- Firm income: including both free cash flows and disbursements to/from stakeholders.
- Firm growth in: sales, expenses, capital, total value, and market value of equity.[4]
- Firm returns: yearly, monthly, daily, both raw and adjusted for Fama-French factors.
- Firm income growth: growth in cash flows and disbursements.[5]
- Firm (net) investment: both total and physical investment net of asset sales.

while the typical candidates in the literature — the Normal, Laplace, and Levy-Stable (aka Pareto-Stable or Power-Law) are generally strongly rejected. In likelihood-based information criteria "horse-race" tests, such as Akaike's Information Criterion (AIC) or the Bayesian Information Criterion (BIC), the DLN is overwhelmingly favored over the other candidates.

[4]With total and equity values adjusted for disbursements.

[5]The methods paper Parham (2022a) shows how to derive growth in DLN variates which may sometimes be negative, thus extending the definition of growth measurement from $\mathbb{R}^+$ to $\mathbb{R}$.



(a) standard DLN, N, LN

(b) Std. DLN w/ corrs

(c) DLN w/ corrs

(d) ADLN w/ corrs

Fig. 1. DLN Examples. Panel (a) graphs the PDFs of the standard Normal, log-Normal, and DLN. Panel (b) graphs the PDFs of standard DLN with different correlation coefficients ρ_{pn}. Panel (c) presents the PDFs of a DLN with parameters $(3, 2, 2, 2)$, common in practice, and varying correlation coefficients ρ_{pn}. Panel (c) presents the PDF for the range ± 10, which is a significant truncation due to the long tails of this DLN. Panel (d) presents the same PDFs as Panel (c), but the x-axis is asinh-transformed, such that it spans the range sinh(-10) $\approx$ -11,000 to sinh(10) $\approx$ 11,000. Formal derivations are in Parham (2022a).

Figure 2 graphically presents several of the firm-related findings from Parham (2022b). Panels (a)-(c) show the distribution of firm cash flows, with Panel (a) showing the untransformed but truncated data, Panel (b) showing the asinh-transformed data, and Panel (c) showing its q-q plot vs. the DLN, with excellent fit. The next six panels exhibit similarly excellent visual fits for: the growth of firm sales, the Fama-French factor-adjusted equity returns at monthly frequency, and net total yearly investment. The fit between the DLN and equity returns is especially noteworthy, given the voluminous literature on the determinants, fat-tails, and statistical properties of equity returns.

Why is the firm so remarkably governed by the DLN distribution, and what causes the heavy tails of firm growth and returns? To formalize the mechanism, Parham (2022c) presents a simple neo-classical model of firm dynamics (aka a "q-theory" model) which replicates these findings. The core innovation introduced into the model is using the most fundamental "sources and uses" equation of the firm: *income = sales - expenses*. The model does that by replacing the neo-classical log-linear, or "Cobb-Douglas", production function

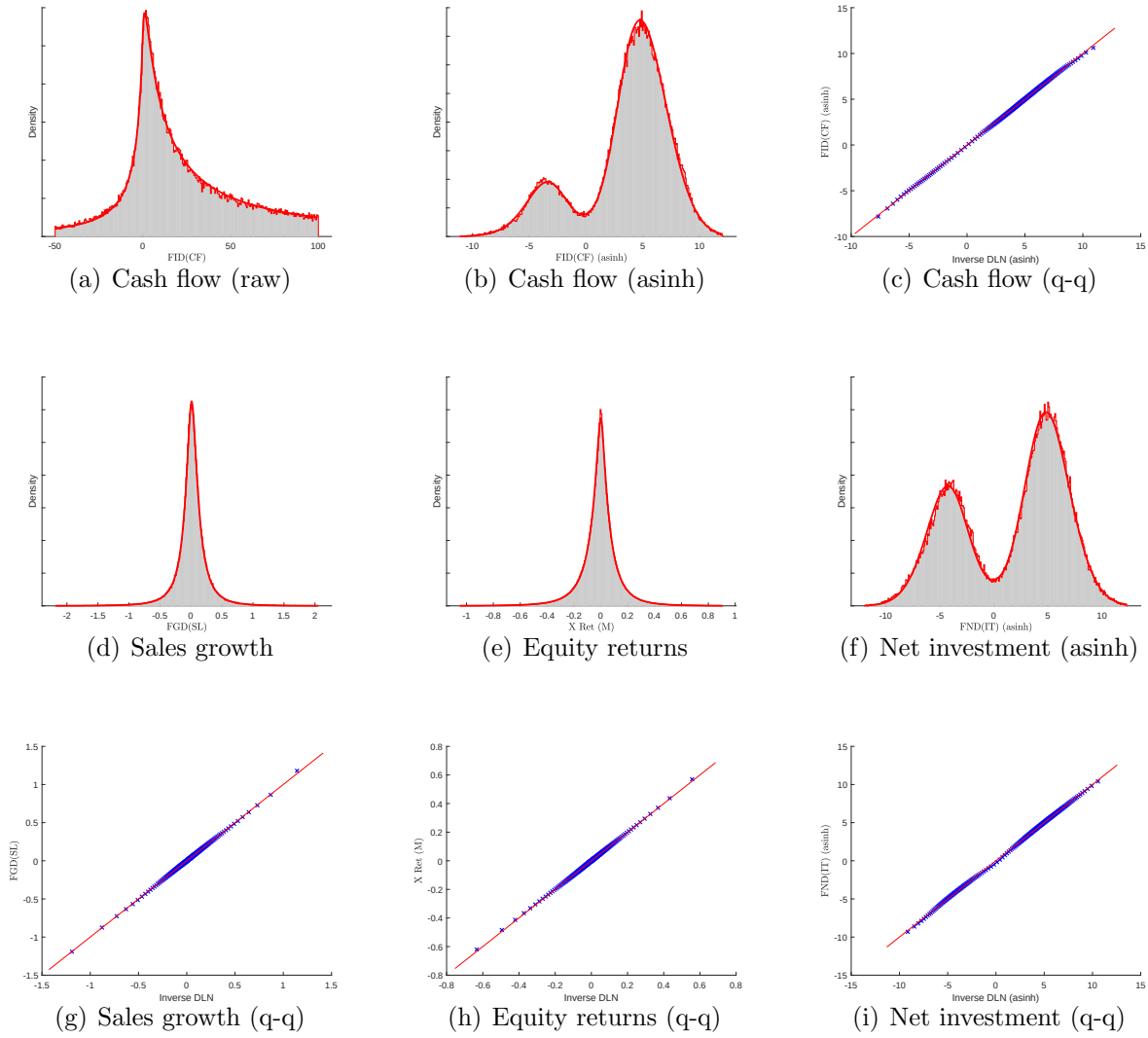
$$\mathbf{Y}_z\left(K_t, Z_t\right) = \underbrace{Z_t \cdot K_t^{\theta_Z}}_{Income} = \exp\left(z_t + \theta_Z \cdot k_t\right) \tag{10}$$

with a difference-of-log-linears production function explicitly modeling sales and expenses

$$\mathbf{Y}_{sx}\left(K_t, S_t, X_t\right) = \underbrace{S_t \cdot K_t^{\theta_S}}_{Sales} - \underbrace{X_t \cdot K_t^{\theta_X}}_{Expenses} = \exp\left(s_t + \theta_S \cdot k_t\right) - \exp\left(x_t + \theta_X \cdot k_t\right) \tag{11}$$

With the stochastic productivity variables Z_t, S_t, X_t following AR(1) in logs with Normal innovations, $\theta_Z, \theta_S, \theta_X$ returns-to-scale coefficients, K_t firm capital, and lower case variables denoting logged values.

The canonical functional form $\mathbf{Y}_z$ used throughout neo-classical economics directly models income, i.e., sales net of expenses. Because $\mathbf{Y}_z \geq 0$, it implicitly assumes that expenses are always lower than sales, and in fact, are a fixed percentage of sales. It also assumes the two scale with the same returns-to-scale coefficient. Parham (2022c) shows that deviating from these assumptions by using the $\mathbf{Y}_{sx}$ functional form is sufficient to replicate the DLN distributions of income, growth, returns, and investment.



(a) Cash flow (raw) (b) Cash flow (asinh) (c) Cash flow (q-q)

(d) Sales growth (e) Equity returns (f) Net investment (asinh)

(g) Sales growth (q-q) (h) Equity returns (q-q) (i) Net investment (q-q)

Fig. 2. Firm distributions stylized facts. Panels (a)-(c) present the distribution of cash flows, with raw but truncated values in (a), asinh-transformed values in (b), and q-q plot vs. the DLN in (c). Panels (d) and (g) present the distribution and q-q vs. DLN for sales growth. Panels (e) and (h) repeat for Fama-French factor-adjusted equity returns at monthly frequency, Panels (f) and (g) repeat for asinh-transformed net investment. Precise variable definitions and further statistical tests are in Parham (2022b).

Returning to population dynamics, Parham (2022d) uses data from the Census's ACS and from the BEA to show that city growth distributes DLN as well. This holds whether one uses counties or MSAs, and whether one measures population growth or local GDP growth. Parham (2022d) then writes a simple model of city dynamics, nearly isomorphic to the firm dynamics model of Parham (2022c) described above, which replicates the size and growth distributions of US locales.

The core feature of the model is again a production function similar to $\mathbf{Y}_{sx}$ from Equation 11 above. The production function models the flow of net economic value created by a city as the interplay between agglomeration benefits and congestion costs. This economic flow is captured by city inhabitants, firms operating in the city, the government, or in general the social planner. Both benefits and costs increase with city size, just as both sales and expenses increase with firm size. But the interplay between them may give rise to positive net economic flow ($Y_{sx} > 0$) leading to immigration into the city, or to negative net economic flow ($Y_{sx} < 0$) leading to emigration out of the city.

Summary

A plethora of phenomena in nature arise as the balance of two opposing forces. When these two forces are themselves multiplicative combinations of underlying latent random forces, the phenomena will tend to distribute DLN. Growth phenomena are especially likely to be DLN, as growth is at its essence a multiplicative (i.e. exponential) process. Hence, the forces supporting growth and the forces opposing it are likely to be log-Normal, and growth itself is likely to be DLN. I hence posit that the DLN is a *fundamental distribution in nature*, in the sense that it emerges in many disparate natural phenomena, similar to the repeated disparate emergence of the Normal and log-Normal distributions.

This paper reviews the results of four companion papers. The first, Parham (2022a), is a methods paper. It formally defines the DLN, derives its PDF and CDF, closed-firm terms for its moments, MLE estimation procedures, etc. Interestingly, Parham (2022a) shows how to extend the notion of growth measurement to growth from negative values, when the variate whose growth we measure distributes DLN, thus extending the definition of growth from being supported on $\mathbb{R}^+$ to $\mathbb{R}$.[6]

The second, Parham (2022b), analyzes data on all public US firms in the 50-year period 1970-2019. It establishes stylized facts regarding firm scale and growth, showing among other facts that the DLN is the core distribution governing firm dynamics. Specifically, it rejects the: (i) Normal, proposed as the governing distribution since Gibrat (1931); (ii) Lévy-Stable, proposed since the early works of Mandelbrot (1960, 1961, 1962, 1963) and Fama (1963, 1965), and still used as the de-facto standard when modeling, e.g., equity returns; (iii) the Laplace, proposed by Stanley, Amaral, Buldyrev, Havlin, Leschhorn, Maass, Salinger, and Stanley (1996) and popular in the "econo-physics" literature. Information-criteria-based "horse-race" tests on more than 40 firm distributions and flavors thereof considered by Parham (2022b) *all* overwhelmingly favor the DLN over the other candidates.

[6] E.g., a firm that had \$100M of profit last year and \$120M of profit this year grew its profit by $\log(120) - \log(100) = 0.182$ log-points. But what if the firm had \$100M of *losses* last year?

The third, Parham (2022c), is a theory paper. It writes down a simple neo-classical model of firm dynamics, in the spirit of Hayashi (1982) and Hennessy and Whited (2005, 2007). The core innovation is in modeling sales and expenses separately, rather than modeling income directly. I.e., it replaces the canonical neo-classical log-linear (or "Cobb-Douglas") production function with a difference-of-log-linears function. Parham (2022c) shows that this suffices to yield simulated firms governed by the DLN distribution, commensurate with the stylized facts in Parham (2022b).

The fourth, Parham (2022d), applies the same reasoning to the problem of modeling city dynamics, in the spirit of Henderson (1974). It shows that city growth distributes DLN as well and writes a model nearly isomorphic to the firm model of Parham (2022c). The model replicates the stylized facts of city size and growth. Interestingly, the city size distribution in this model and the firm size distribution in the firm model are both log-skewed-Normal (i.e., firm and city *scale*, defined as the log of size, are skewed-Normal). This is commensurate with the data on the entire (rather than just tail) distribution of both firms and cities.

The DLN can hence be used both as a statistical model for a variety of phenomena and as a guiding principle in developing theories regarding the mechanisms underlying these phenomena.

References

Fama, E. F., 1963. Mandelbrot and the Stable Paretian Hypothesis. The Journal of Business 36, 420–429.

Fama, E. F., 1965. The Behavior of Stock-Market Prices. The Journal of Business 38, 34–105.

Gibrat, R., 1931. Les Inégalités Économiques. Librairie du Recueil Sirey, Paris.

Gulisashvili, A., Tankov, P., 2016. Tail behavior of sums and differences of log-normal random variables. Bernoulli 22, 444–493.

Hayashi, F., 1982. Tobin's Marginal q and Average q: A Neoclassical Interpretation. Econometrica 50, 213–224.

Henderson, J. V., 1974. The Sizes and Types of Cities. The American Economic Review 64, 640–656.

Hennessy, C. A., Whited, T. M., 2005. Debt Dynamics. The Journal of Finance 60, 1129–1165.

Hennessy, C. A., Whited, T. M., 2007. How Costly Is External Financing? Evidence from a Structural Estimation. The Journal of Finance 62, 1705–1745.

Lo, C. F., 2012. The Sum and Difference of Two Lognormal Random Variables. Journal of Applied Mathematics 2012, 1–13.

Malthus, T., 1798. An Essay on the Principle of Population. J. Johnson, London.

Mandelbrot, B., 1960. The Pareto-Lévy Law and the Distribution of Income. International Economic Review 1, 79–106.

Mandelbrot, B., 1961. Stable Paretian Random Functions and the Multiplicative Variation of Income. Econometrica 29, 517–543.

Mandelbrot, B., 1962. Paretian Distributions and Income Maximization. The Quarterly Journal of Economics 76, 57–85.

Mandelbrot, B., 1963. New Methods in Statistical Economics. Journal of Political Economy 71, 421–440.

Parham, R., 2022a. 1. The Difference-of-Log-Normals Distribution: Properties, Estimation, and Growth. Mimeo .

Parham, R., 2022b. 2. Facts of US Firm Scale and Growth 1970-2019: An Illustrated Guide. Mimeo .

Parham, R., 2022c. 3. Why Are Firm Growth Distributions Heavy-Tailed? Mimeo .

Parham, R., 2022d. 4. The Scale and Growth Distributions of Cities. Mimeo .

Stanley, M. H. R., Amaral, L. A. N., Buldyrev, S. V., Havlin, S., Leschhorn, H., Maass, P., Salinger, M. A., Stanley, H. E., 1996. Scaling behaviour in the growth of companies. Nature 379, 804.