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

Robert Parham*

Saturday 11th February, 2023

Summary

The growth of many natural and social phenomena including pandemics[1], firms,[2] cities,[3] and various economic indices,[4,5] is known to be heavy-tailed. Most growth is modest, but we often observe explosive growth rates, such as firms doubling or halving in size within a short period. Neither a simple explanation nor a well-fitting distributional form for these growth phenomena is known. Here we show that a hitherto obscure statistical distribution — the Difference-of-Log-Normals (DLN) — describes a plethora of growth phenomena remarkably well, and discuss why it arises as a natural consequence of the Central Limit Theorem (CLT). Our results demonstrate how growth phenomena subject to opposing random exponential forces are likely to distribute DLN. This provides both a framework for scientifically modeling these phenomena and a simple distributional form to be used when empirically modeling observed heavy-tailed growth. We hence posit that the DLN is a *fundamental distribution in nature*, in the sense that it emerges in many disparate natural phenomena, especially growth phenomena, similar to the repeated disparate emergence of the Normal and log-Normal distributions.

*University of Virginia (robertp@virginia.edu).

What do the growth rates of such varied phenomena as new COVID-19 infection cases, tests conducted, and vaccinations administered; firm sales, capital, income, and stock values; and city populations and regional GDPs have in common? They all appear to distribute as the Difference-of-Log-Normals (Figures 1- 3).

The Difference-of-Log-Normals distribution, henceforth DLN, 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, i.e. $W \sim \mathrm{DLN}(\mu_p, \sigma_p, \mu_n, \sigma_n, \rho_{pn})$, and fully derive its properties elsewhere.[6]

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, we were unable to find instances of using it anywhere in the sciences, and only two statistical works considering it.[7,8] 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, we posit that the DLN is a *fundamental distribution in nature*, likely describing a plethora of natural and economic phenomena.[9]

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",[10] 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, epidemic spread, 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 the natural phenomenon will tend to distribute 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.

To fix ideas, Figure 4 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,[11] 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.

Possibly the most intuitive example of the DLN's emergence is in the context of a simple population dynamics ("birth-death") model.[12] 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:

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

in which $b\left(t\right) \geq 0$ and $d\left(t\right) \geq 0$ are the instantaneous birth and death rates. Generally, $b\left(t\right)$ and $d\left(t\right)$ are stochastic, depending on some underlying latent forces such as food availability, climate, predation, etc. Because negative birth or death rates are inadmissible, we cannot assume they are jointly Normal. The next-simplest hypothesis (in the maximum entropy sense) for their distribution is hence the bi-variate log-Normal. This means the distribution of their difference, or the distribution of population growth in the model, is DLN — providing an intuitive explanation to the emergence of DLN in the COVID-19 data, described in Figure 1.

Moving on to the realm of finance, consider the most fundamental "sources and uses" equation of the firm: *income = sales - expenses*. Both sales and expenses are approximately log-Normally distributed, and it is standard practice in neo-classical economics to model income as a controlled AR(1) stochastic Markov process in logs, with Normal innovations. In such models, growth is counter-factually Normally distributed. If we instead model sales and expenses *separately* as two co-controlled AR(1) stochastic Markov processes in logs, with (possibly correlated) Normal innovations, the model then predicts firm income (and consequently, growth) will distribute DLN. In essence, we replace the neo-classical log-linear, or "Cobb-Douglas", production function

$$\mathbf{Y}_z\left(K_t, Z_t\right) = \underbrace{Z_t \cdot K_t^{\theta_Z}}_{Income} = \exp\left(z_t + \theta_Z \cdot k_t\right) \tag{8}$$

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

$$\mathbf{Y}_{sx}\left(K_t, S_t, X_t\right) = \underbrace{S_t \cdot K_t^{\theta_S}}_{Sales \equiv \mathbb{S}_t} - \underbrace{X_t \cdot K_t^{\theta_X}}_{Expenses \equiv \mathbb{X}_t} = \exp\left(s_t + \theta_S \cdot k_t\right) - \exp\left(x_t + \theta_X \cdot k_t\right) \tag{9}$$

With the stochastic log-productivity variables z_t, s_t, s_t following AR(1) with Normal innovations, $\theta_Z, \theta_S, \theta_X$ returns-to-scale coefficients, and k_t logged firm capital. Importantly, Equation 9 can be decomposed such that

$$\mathbf{Y}_{sx}\left(K_t, S_t, X_t\right) = 2 \cdot \exp\left(\lambda_t\right) \cdot \sinh\left(\tau_t\right)$$

$$\lambda_t = \lambda(k_t, s_t, x_t) = \frac{s_t + x_t}{2} + \frac{\theta_S + \theta_X}{2} \cdot k_t = \log(\sqrt{\mathbb{S}_t \cdot \mathbb{X}_t}) \tag{10}$$

$$\tau_t = \tau(k_t, s_t, x_t) = \frac{s_t - x_t}{2} + \frac{\theta_S - \theta_X}{2} \cdot k_t = \log(\sqrt{\mathbb{S}_t / \mathbb{X}_t})$$

with λ_t firm *scale*, and τ_t firm *efficiency*, thus justifying our use of the asinh transform.

To test these predictions of the model, we empirically analyze the data on all public US firms in the 50-year period 1970-2019. Figure 2 graphically presents several of our findings. Panels (a)-(c) show the distribution of firm income, with Panel (a) showing the un-transformed 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.

In a battery of tests, the DLN is shown to be the core distribution driving firm dynamics.[11] 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.
- Firm returns: yearly, monthly, daily, both raw and adjusted for Fama-French factors.
- Firm income growth: growth in cash flows and disbursements.
- Firm (net) investment: both total and physical investment net of asset sales.

while the typical candidates in the literature — the Normal, Laplace, and Levý-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.

Returning to population dynamics, Figure 3 presents data on population growth,[13] and on economic activity growth by county and by metropolitan area and industry from the US Bureau of Economic Analysis. The DLN arises again, as can be seen visually in the q-q plots. A "horse-race" with the other typical candidate distributions again strongly favors the DLN. This finding is a natural outcome of a simple model of city dynamics[14] in which cities grow subject to the interplay between agglomeration benefits and congestion costs, both of which exert exponential influence on the flow of aggregate economic value created by cities. 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 (i.e., $Y_{sx} > 0$) leading to immigration into the city, or to negative net economic flow (i.e., $Y_{sx} < 0$) leading to emigration out of the city.

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 distribute DLN. This insight is useful both when constructing models to describe these phenomena, as briefly outlined above for firms and cities, and when empirically modeling such phenomena by providing a simple distributional form to estimate and use.

References

[1] Kris V. Parag, Christl A. Donnelly, and Alexander E. Zarebski. "Quantifying the Information in Noisy Epidemic Curves". In: *Nature Computational Science* 2.9 (Sept. 2022), pp. 584–594. ISSN: 2662-8457. DOI: 10.1038/s43588-022-00313-1.

[2] T. S. Ashton. "The Growth of Textile Businesses in the Oldham District, 1884-1924". In: *Journal of the Royal Statistical Society* 89.3 (1926), pp. 567–583. ISSN: 0952-8385. DOI: 10.2307/2341720.

[3] Jan Eeckhout. "Gibrat's Law for (All) Cities". In: *The American Economic Review* 94.5 (2004), pp. 1429–1451. ISSN: 0002-8282.

[4] Rosario N. Mantegna and H. Eugene Stanley. "Scaling Behaviour in the Dynamics of an Economic Index". In: *Nature* 376.6535 (July 1995), p. 46. ISSN: 1476-4687. DOI: 10.1038/376046a0.

[5] Michael H. R. Stanley et al. "Scaling Behaviour in the Growth of Companies". In: *Nature* 379.6568 (Feb. 1996), p. 804. ISSN: 1476-4687. DOI: 10.1038/379804a0.

[6] Robert Parham. "The Difference-of-Log-Normals Distribution: Properties, Estimation, and Growth". In: *arXiv:2302.02486 [stat.ME]* (2023). arXiv: 2302.02486 [stat.ME].

[7] C. F. Lo. "The Sum and Difference of Two Lognormal Random Variables". In: *Journal of Applied Mathematics* 2012 (2012), pp. 1–13. ISSN: 1110-757X, 1687-0042. DOI: 10.1155/2012/838397.

[8] Archil Gulisashvili and Peter Tankov. "Tail Behavior of Sums and Differences of Log-Normal Random Variables". In: *Bernoulli* 22.1 (Feb. 2016), pp. 444–493. ISSN: 1350-7265. DOI: 10.3150/14-BEJ665. arXiv: 1309.3057 [math, q-fin].

[9] Geoffrey West. *Scale: The Universal Laws of Growth, Innovation, Sustainability, and the Pace of Life in Organisms, Cities, Economies, and Companies*. First Edition. New York: Penguin Press, May 2017. ISBN: 978-1-59420-558-3.

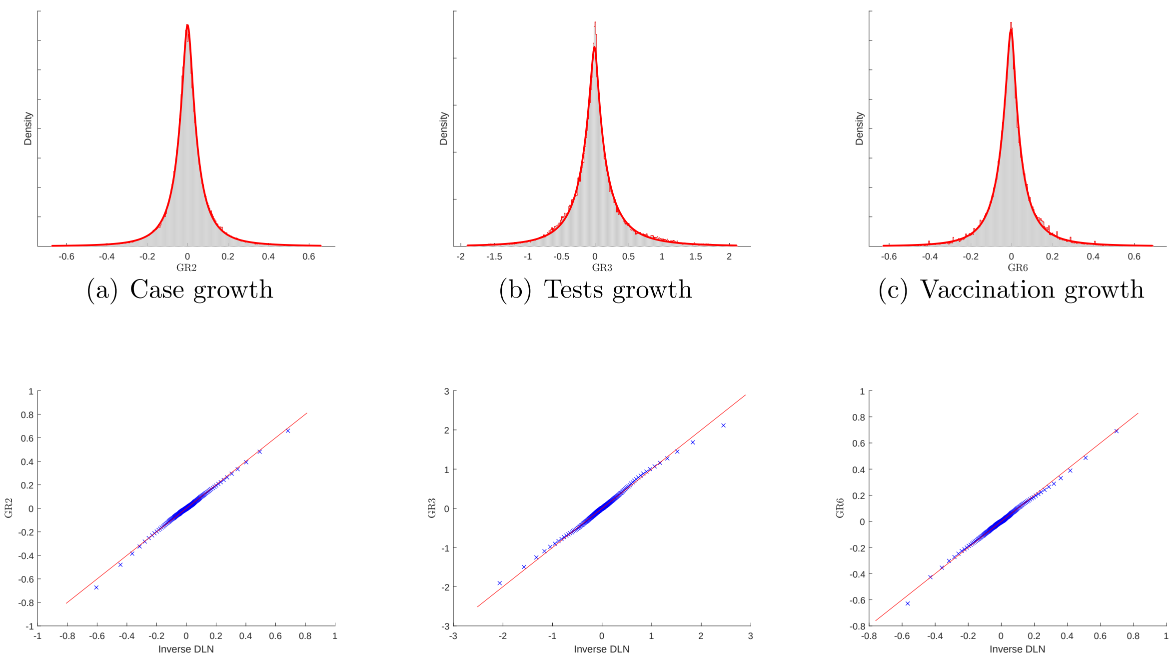
[10] R Gibrat. *Les Inégalités Économiques.* Paris: Librairie du Recueil Sirey, 1931.

[11] Robert Parham. "Facts of US Firm Scale and Growth 1970-2019: An Illustrated Guide". In: *arXiv:2302.02485 [q-fin.GN]* (2023). arXiv: 2302.02485 [q-fin.GN].

[12] Thomas Malthus. *An Essay on the Principle of Population.* London: J. Johnson, 1798.

[13] Hernán D Rozenfeld et al. "The Area and Population of Cities: New Insights from a Different Perspective on Cities". In: *American Economic Review* 101.5 (Aug. 2011), pp. 2205–2225. ISSN: 0002-8282. DOI: 10.1257/aer.101.5.2205.

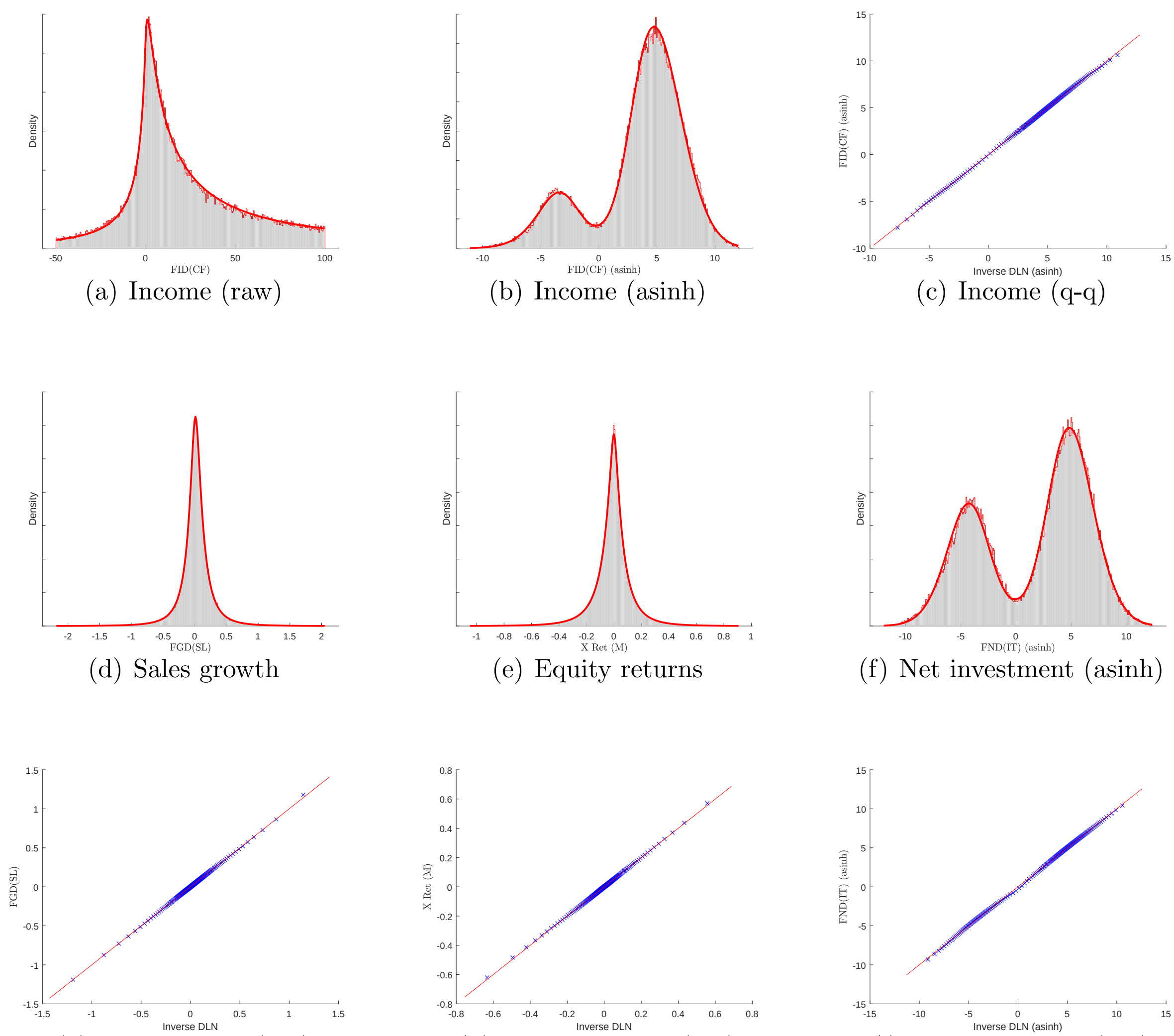
[14] J. V. Henderson. "The Sizes and Types of Cities". In: *The American Economic Review* 64.4 (1974), pp. 640–656. ISSN: 0002-8282.



(a) Case growth (b) Tests growth (c) Vaccination growth

(d) Case growth (q-q) (e) Tests growth (q-q) (f) Vacc. growth (q-q)

Fig. 1. Covid distributions stylized facts. Panels (a)-(c) present the distributions of daily new cases growth, new tests conducted growth, and new vaccinations growth, respectively. Panels (d)-(f) present the respective q-q plots.

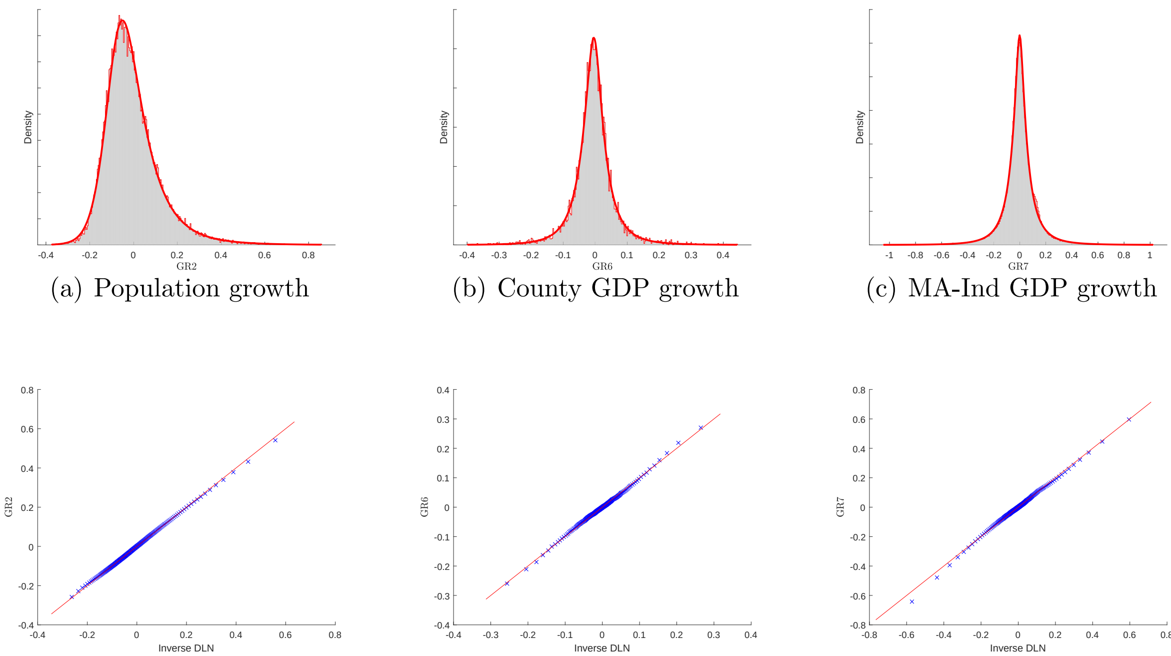


(a) Income (raw) (b) Income (asinh) (c) Income (q-q)

(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)

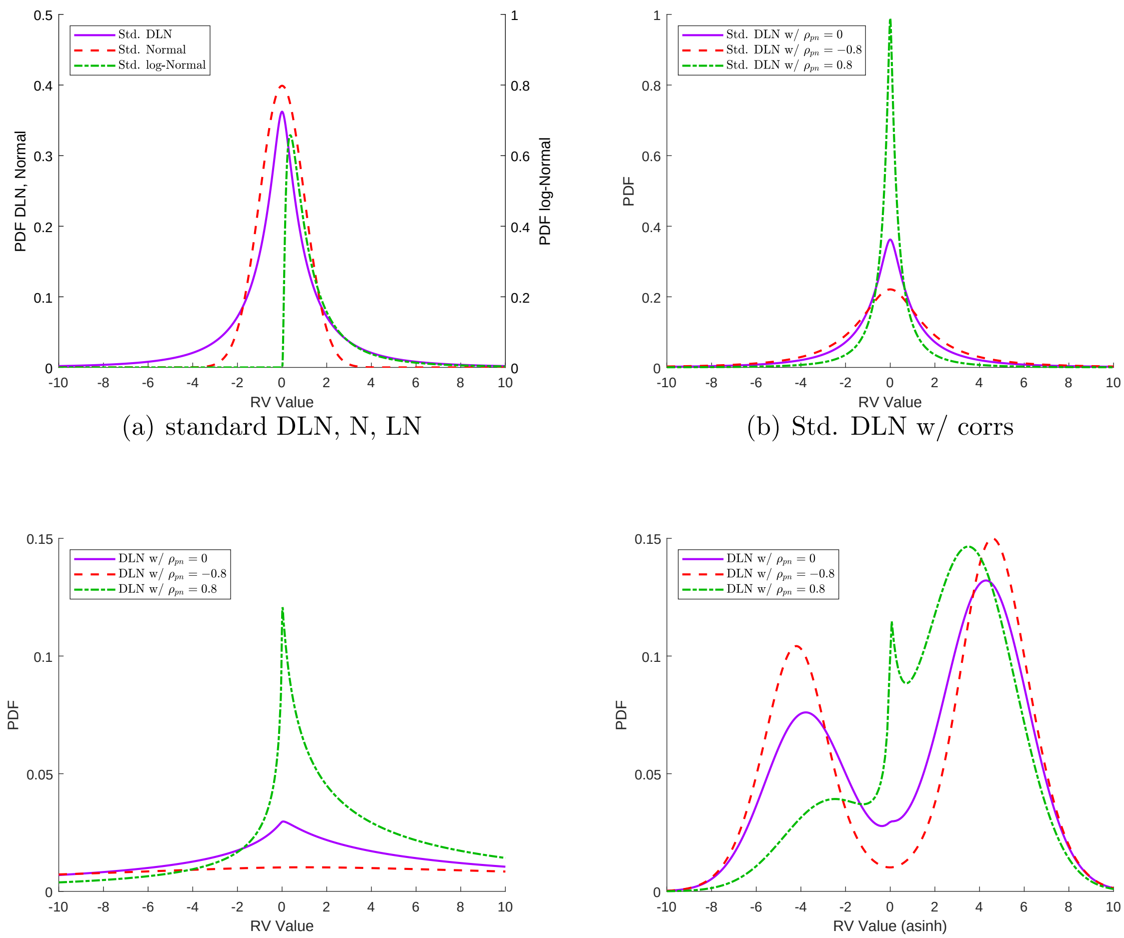
Fig. 2. Firm distributions stylized facts. Panels (a)-(c) present the distribution of income, with raw but truncated values in (a), asinh-transformed values in (b), and q-q plot vs. the DLN in (c). Panels (d) and (g) present the distribution and q-q vs. DLN for sales growth. Panels (e) and (h) repeat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



(a) Population growth (b) County GDP growth (c) MA-Ind GDP growth

(d) Pop growth (q-q) (e) GDP growth (q-q) (f) GDP growth (q-q)

Fig. 3. City distributions stylized facts. Panels (a)-(c) present the distributions of population growth, GDP growth by county, and GDP growth by metropolitan area and industry, respectively. Panels (d)-(f) present the respective q-q plots.



(a) standard DLN, N, LN

(b) Std. DLN w/ corrs

(c) DLN w/ corrs

(d) ADLN w/ corrs

Fig. 4. 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.

A Methods

A.1 Data

COVID-19 data analyzed in Figure 1 are from Our World In Data (ourworldindata.org/coronavirus). We use daily worldwide data, and plot all available growth observations when the base value is higher than 10 (e.g., more than 10 infections per day or more than 10 vaccines given), and the growth rate is different from 0 (as growth being exactly 0 usually indicates stale data). The data were downloaded on 2/5/2022 and cover 143K observations for case growth, 71K observations for tests growth, and 112K observations for vaccination growth.

Firm data analyzed in Figure 2 are from the Compustat/CRSP-combined dataset, accessed via Wharton's WRDS. The data cover 164K firm-year observations on 15,797 firms between 1970-2019. Sample selection criteria, exact variable definitions, and descriptive statistics are reported in [11].

City data presented in 3 are from three sources. The data in Panels (a) and (d) are from [13], and pertain to population growth from 1991 to 2000 in 46K locales identified by the clustering algorithm of [13]. The data in Panels (b) and (e) are from the US Bureau for Economic Analysis (Gross Domestic Product by County, 2017-2020), and pertain to 9,333 county-year observations on the growth of per-county GDP for 3,111 counties from 2017 to 2020. The data in Panels (c) and (f) are again from the US BEA (Gross Domestic Product by Metropolitan Area and Industry, 2001-2017), and pertain to 270K growth observations on 87 industries within 384 MAs over 17 years.

A.2 Analysis

For each growth measure, the data are first fit to the DLN distribution using the MLE estimator described in [6]. The empirical distribution of the data, along with the fitted DLN (in red) are presented as figures, along with a q-q plot of the empirical CDF vs. the theoretical DLN CDF, also developed in [6].

Next, three statistical goodness-of-fit tests are used to verify whether the empirical data indeed stem from the DLN distribution. The three tests used are the Kolmogorov-Smirnov test, the Anderson-Darling test, and the Chi-square test. Details on conducting these tests and on constructing the p-values for the tests are available in [6] and [11]. The tests generally do not reject the DLN for the growth data at the 5% confidence level. This is in contrast with the two other candidate distributions discussed in the literature: the Laplace distribution (itself a difference of exponentially distributed variates), and the Leúy-Stable (aka Pareto-Stable or Power-Law) distribution. Both of these distributions are generally rejected at the 5% level by the goodness-of-fit tests.

Finally, for each growth measure, we conduct likelihood-based information criteria "horse-race" tests between the DLN and the two other candidate distributions. The tests are based on Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC). In all such tests, the DLN is overwhelmingly favored over the other candidates.

B Data Availability

COVID data are freely and publicly available from Our World In Data (ourworldindata.org/coronavirus).

Firm data are publicly available from the CRSP/Compustat Merged Database, which is subscription-fee-based (crsp.org). The firm data were normalized by the year's nominal GDP from the St. Louis Fed (fred.stlouisfed.org).

City data are freely and publicly available from two sources: the US Bureau for Economic Analysis (bea.gov) and a replication package for [13] from the AER's website (aeaweb.org).

C Code Availability

Code to reproduce all figures reported in this analysis, including code implementing the CDF, PDF, and DLN parameter estimation is publicly available from the author and will be provided to Nature.