Vol. 16(2), pp. 10-25, April-June 2024
DOI: 10.5897/IJPC2024.0702
Article Number: 29E9DE572216
ISSN 2141-2499
Copyright © 2024
Author(s) retain the copyright of this article
http://www.academicjournals.org/IJPC

International Journal of Psychology and Counselling

Full Length Research Paper

Lognormal distribution for social researchers: A probability classic

José Moral de la Rubia

School of Psychology, Facultad de Psicología, Universidad Autónoma de Nuevo León, Monterrey, Nuevo León, México.

Received 5 February, 2024; Accepted 31 May, 2024

This academic article aims to present the lognormal distribution clearly, accompanied by an example applied to sexual behavior, facilitating understanding among social researchers. This distribution, characterized by positive skewness, thin shoulders, and heavy tails, serves as a robust probability model for various social and behavioral variables. It is developed from its two-parameter format, a location parameter (μ) and a squared scale parameter (σ²). The paper begins with a historical note on the relationship of the lognormal distribution to the normal distribution. The density, cumulative, and characteristic functions of the distribution are shown. Although it has an analytical expression for the nth order moment, it is not determined by its moments, lacking a moment generating function. Following the presentation of these functions, measures of central tendency, variability, and shape are discussed. The estimators of μ and σ² using the methods of moments and maximum likelihood are then introduced. Some of their mathematical properties and the calculation of dispersion intervals for 68.3, 95.4 and 99.7% of the data are presented. All this material is applied to two examples of probability calculation, descriptive measures, and parameter estimation related to sexual behavior. Finally, suggestions are provided for the practical application of the lognormal distribution.

Key words: Probability distribution, continuous variable, parameter estimation, arithmetic descriptive measures, geometric descriptive measures.

INTRODUCTION

The purpose of this academic paper is to disseminate statistical knowledge about a probability distribution and to facilitate its practical application. Specifically, it focuses on the lognormal distribution, which is one of the most studied non-normal continuous distributions in statistics and probability theory, and can be highly useful in the field of social sciences. However, it is relatively unknown among researchers in these disciplines, except in the realms of economic and business sciences (Minyu et al., 2020). It is important to note that the explanatory exposition of this distribution is often very theoretical (Al-Masri, 2022) and can be confusing, particularly because it is presented under different parameterizations that may be intermixed in the same publication (Swat et al., 2016; Wikipedia Contributors, 2023). Therefore, the objective of this article is to present this probability distribution in a clear, understandable, and exemplified manner, with an example applied to the field of sexual behavior. The data

E-mail: jose_moral@hotmail.com. or jose.morald@uanl.edu.mx. Tel: (00 52 81) 8333 8233.

for the example were generated or simulated. It should be clarified that this is not an empirical research report.

The lognormal distribution has its roots in a study conducted by the English biologist and mathematician Francis Galton in the year 1879 on the distribution of the geometric mean. This work was further explored by the Scottish physician Donald McAlister in the year 1879, leading to the alternative name Galton's distribution (Rao et al., 2022). The Dutch astronomer Jacobus Cornelius Kapteyn in the year 1903 was among the first to apply this distribution to the examination of biological variables. The lognormal distribution is characterized by support in the positive real numbers, positive skewness, thin shoulders, heavy tail (leptokurtosis), a large variance, and a median corresponding to its geometric mean.

This distribution is applicable to various measurements in living tissues, such as the weight and length of skin, as well as the length of inert appendages like hair, claws, nails, and teeth. Additionally, it finds relevance in modeling the spread of epidemics, including its recent use in representing the spread of COVID-19 (Kapteyn and van Uven, 1916; Lahcene, 2021; Rao et al., 2022). Cobb and Douglas (1928) introduced a function as a probability model for production in a country, aligning with the lognormal distribution. This economic application has seen widespread use, although it is presently undergoing revision (Gechert et al., 2019; Smirnov and Wang, 2021). In the realm of sexual behavior, the lognormal distribution is employed in studying various aspects. Examples include analyzing the number of sexual partners (Gualandi and Toscani, 2019; Major et al., 2021), examining networks of sexual contacts (Ito et al., 2022), and studying the frequency of forced sexual acts (Akampurira, 2022).

CHARACTERIZATION OF THE DISTRIBUTION

It begins with the exposition of the different functions, parameters, and the probability pattern of this continuous distribution. The lognormal distribution is characterized by two parameters in logarithmic scale: a location parameter μ = $\mu_{\ln(X)}$ with a parameter space of (–∞, +∞), and a squared scale parameter σ^2 = $\sigma^2_{\ln(X)}$ with a parameter space of (0, +∞). It is denoted as *Lognormal* (μ, σ^2). The e-based exponential of the location parameter μ provides the geometric mean and median of the distribution. These two measures are considered more appropriate indicators of central tendency than the arithmetic mean for a distribution with positive skewness and leptokurtosis.

Holding the squared scale parameter σ^2 constant, the peak of the distribution (modal value or maximum density) decreases as μ increases (Figure 1). The e-based exponential of the squared scale parameter σ^2 provides the geometric variance, and the unsquared value yields the geometric standard deviation σ, which are considered superior measures of variability compared to the arithmetic variance and standard deviation for this distribution.

Holding the location parameter μ constant, the shoulder area thins, and the tail area thickens as σ^2 becomes larger (Figure 2). Additionally, a three-parameter presentation introduces a threshold parameter *a*, representing the minimum value. This ensures that the origin of the distribution is not at 0 but at a value other than 0, and it can be negative: x ≥ a ∊ X ~ *Lognormal* (*a*, μ, σ^2), where a ∊ R.

The lognormal distribution can be obtained from the normal distribution (Galton, 1879; McAlister, 1879). If a random variable Y follows a normal distribution with a location parameter μ and a squared scale parameter σ², the exponential function with base e and exponent Y follows a lognormal distribution with a location parameter μ and a squared scale parameter σ²: Y ~ *N* (μ, σ²) ⟹ X = e^Y ~ *Lognormal* (μ, σ²). Conversely, if a random variable X follows a lognormal distribution with a location parameter μ and a squared scale parameter σ², its natural logarithm follows a normal distribution with a location parameter μ and a squared scale parameter σ²: X ~ *Lognormal* (μ, σ²) ⟹ Y = ln(X) ~ *N* (μ, σ²). It should be noted that the natural basis is the one usually employed, but any other positive basis is equally useful.

$$X \sim Lognormal(\mu, \sigma^2)$$

Location parameter: $\mu = \mu_{\ln(X)} \in \mathbb{R}$

Squared scale parameter: $\sigma^2 = \sigma^2_{\ln(X)} > 0 \in \mathbb{R}^+$

Support: $x \in (0, \infty)$

Probability density function:

$$f_X(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln(x_i)-\mu)^2}{2\sigma^2}} = \frac{\varphi\left(\frac{\ln(x)-\mu}{\sigma}\right)}{x} = \frac{\varphi\left(\frac{\ln(x)-\mu_{\ln(x)}}{\sigma_{\ln(x)}}\right)}{x}$$

where φ(z) = density function of the standard normal distribution *N* (0, 1).

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

Cumulative distribution function:

$$F_X(x) = P(X \leq x) = \int_0^x f_X(x) d_x = \frac{1}{2}\left[1 + erf\left(\frac{\ln(x_i) - \mu}{\sigma\sqrt{2}}\right)\right] = \Phi\left(\frac{\ln(x_i) - \mu}{\sigma}\right)$$

where *erf* = Gaussian error function,

$$erf(x) = \frac{2}{\sqrt{\pi}} \int_0^\infty e^{-t^2}\, dt = \frac{2}{\sqrt{\pi}} \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{(2n+1) \times n!}; n! = \prod_{i=0}^{n-1} (n - i)$$

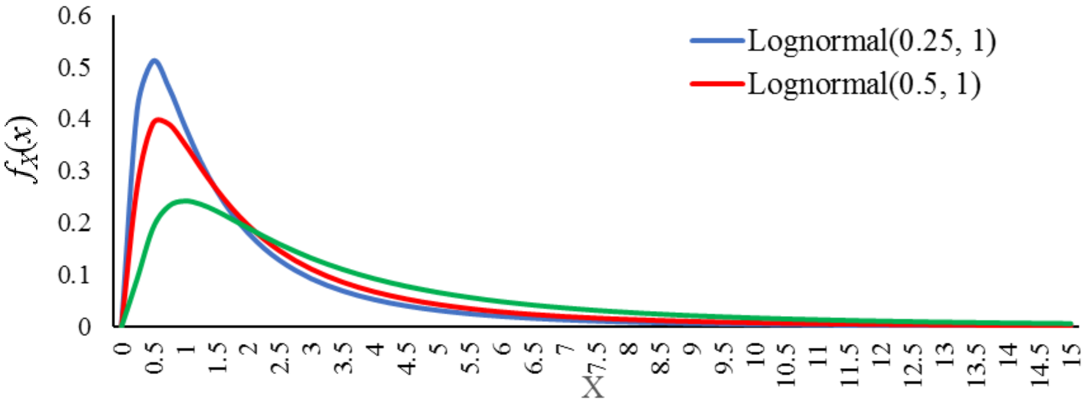

Figure 1. Density functions of three variables with lognormal distribution with location parameters μ = 0.25, 0.5, and 1, and squared scale parameter σ^2 = 1.
Source: Author's elaboration.

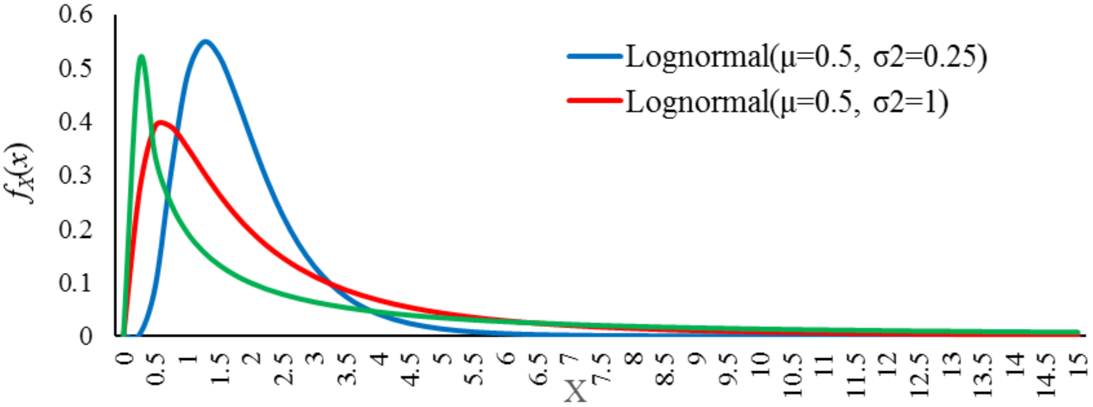

Figure 2. Density functions of three variables with lognormal distribution with location parameter μ = 0.5 and squared scale parameters σ^2 = 0.25, 1, and 4.
Source: Author's elaboration.

where Φ = cumulative distribution function of the standard normal distribution *N* (0, 1).

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$

The quantile function:

$$Q_X(p) = x_{(p)} = e^{\mu + \sigma \times \sqrt{2} \times erf^{-1}(2p-1)} = e^{\mu + \sigma \times \Phi^{-1}(2p-1)}; \; F_X(x) = p$$

where $\Phi^{-1}(p)$ = probit function, *p*-order quantile function or inverse of the cumulative distribution function of the standard normal distribution *N*(0, 1).

Although the lognormal distribution has all its non-central moments defined, this distribution is not determined by its moments, so a moment generating function, $M_X(t)$, cannot be defined around a non-trivial value of *t* in the set of real numbers. However, the characteristic function, $C_X(t)$, can be defined for real values of *t* (origin) that yield complex values with a positive imaginary component as their image.

$$C_X(t): \mathbb{R} \longrightarrow \mathbb{C}; C_X(t) = E\left(e^{itx}\right) = \int_0^{+\infty} e^{itx} f_X(x) d_x = \sum_{n=0}^{\infty} \frac{(it)^n}{n!} e^{n\mu + \frac{n^2\sigma^2}{2}}$$

$$i = \sqrt{-1} \; and \; t \in R$$

MEASURES OF CENTRAL TENDENCY

Arithmetic mean or mathematical expectation:

$$\mu(X) = E(X) = \int_0^{+\infty} x f_X(x) d_x = e^{\mu + \sigma^2/2} = e^{\mu} e^{\sigma^2/2} = \mu_g(X) \sqrt{\sigma_g^2(X)}$$

The arithmetic mean μ(X) may also be expressed as the product of the geometric mean $\mu_g(X)$ and the square root of the geometric variance $\sqrt{[\sigma_g^2(X)]}$.

Median (*Mdn*):

$$Mdn(X) = Q_X\left(p = \frac{1}{2}\right) = x_{\left(p=\frac{1}{2}\right)} = e^{\mu + \sigma\sqrt{2}erf^{-1}\left(\frac{2}{2}-1\right)} = e^{\mu+0} = e^{\mu}$$

Geometric mean (G):

$$G(X) = \mu_g(X) = e^{E[\ln(X)]} = e^{\mu} = Mdn(X)$$

From the geometric mean, the value of the location parameter μ can be deduced, which is the arithmetic mean or mathematical expectation of the variable transformed logarithmically (with natural base).

$$\mu = \ln[G(X)] = \ln\left[e^{E[\ln(X)]}\right] = E[\ln(X)] = \mu_{\ln(X)}$$

Mode (*Mo*):

$$Mo(X) = e^{\mu-\sigma^2} = e^{\mu}/e^{\sigma^2} = \mu_g(X)/\sigma_g^2(X)$$

The mode *Mo*(X) can be expressed as the ratio of the geometric mean $\mu_g(X)$ to the geometric variance $\sigma_g^2(X)$.

Harmonic mean:

$$H(X) = \mu_h(X) = e^{\mu-\sigma^2/2} = e^{\mu}/\sqrt{e^{\sigma^2}} = \mu_g(X)/\sqrt{\sigma_g^2(X)}$$

The harmonic mean $\mu_h(X)$ can be expressed as the quotient of the geometric mean $\mu_g(X)$ and the square root of the geometric variance $\sqrt{[\sigma_g^2(X)]}$.

The nth-moment:

$$\mu = E(X^n) = \int_0^{+\infty} x^n f_X(x) d_x = e^{n\mu+\frac{n^2\sigma^2}{2}}; n \in \mathbb{Z}^+$$

MEASURES OF VARIABILITY

Arithmetic variance (*Var*), standard deviation (*SD*), and coefficient of variation (*CV*):

$$Var(X) = \sigma^2(X) = \int_0^{+\infty} (x - E(X))^2 f_X(x) d_x = e^{2\mu+\sigma^2}\left(e^{\sigma^2} - 1\right)$$

$$\sigma^2(X) = E(X^2) - E^2(X) = e^{2\mu+2\sigma^2} - \left(e^{\mu+\frac{\sigma^2}{2}}\right)^2 = e^{2\mu+2\sigma^2} - e^{2\mu+\sigma^2} = e^{2\mu+\sigma^2}\left(e^{\sigma^2} - 1\right)$$

$$SD(X) = \sigma(X) = \sqrt{\sigma^2(X)} = \sqrt{e^{2\mu+\sigma^2}\left(e^{\sigma^2} - 1\right)} = e^{\mu+\frac{\sigma^2}{2}}\sqrt{e^{\sigma^2} - 1} = E(X) \times CV(X)$$

$$CV(X) = \frac{SD(X)}{|E(X)|} = \frac{\sigma(X)}{|\mu(X)|} = \frac{e^{\mu+\frac{\sigma^2}{2}}\sqrt{e^{\sigma^2} - 1}}{e^{\mu+\frac{\sigma^2}{2}}} = \sqrt{e^{\sigma^2} - 1}$$

Mean absolute deviation (*MAD*):

$$MAD(X) = E(|X - E(X)|) = 2 \times e^{\mu+\frac{\sigma^2}{2}} \times erf\left(\frac{\sigma}{2\sqrt{2}}\right)$$

where *erf* = Gaussian error function.

Geometric variance (*GV*):

$$GV(X) = \sigma_g^2(X) = E(\ln(X) - E[\ln(X)]) = e^{\sigma^2_{\ln(X)}} = e^{\sigma^2}$$

From the geometric variance, we can derive the value of the squared scale parameter σ^2 which is the variance of the variable transformed logarithmically (with natural base).

$$\sigma^2 = \ln[GV(X)] = \ln[E(\ln(X) - E[\ln(X)])] = \sigma^2_{\ln(X)}$$

Geometric standard deviation (*GSD*):

$$GSD(X) = \sigma = \ln[SD(X)] = \ln\left[\sqrt{E(\ln(X) - E[\ln(X)])}\right] = \sigma_{\ln(X)}$$

It should be noted that geometric standard deviation $\sigma_{\ln(X)}$ and the root of the geometric variance $\sqrt{[\sigma^2_{\ln(X)}]}$ do not constitute an equality.

$$\sigma_{\ln(X)} = e^{\sigma_{\ln(X)}} \neq \sqrt{\sigma^2_{\ln(X)}} = \sqrt{e^{\sigma^2_{\ln(X)}}} = e^{\frac{\sigma^2_{\ln(X)}}{2}}$$

From the geometric standard deviation, we can derive the value of the scale parameter σ, which is the standard deviation of the variable transformed logarithmically (with natural base).

$$\sigma = \ln[SD(\ln(X))] = \ln\left[\sqrt{\ln[E(\ln(X) - E[\ln(X)])]}\right] = \sigma_{\ln(X)}$$

Coefficient of geometric variation (*CGV*) of Kirkwood (1979):

$$CGV(X) = \sigma_g(X) - 1 = GSD(X) - 1 = e^{\sigma} - 1 = e^{\sigma_{\ln(X)}} - 1$$

Shannon's Entropy (H):

$$\mathrm{H}(X) = E(-\log_2[f_X(x)]) = -\int_0^{+\infty} \log_2[f_X(x)]\, f_X(x) d_x = \log_2\left(\sigma e^{\mu+\frac{1}{2}}\sqrt{2\pi}\right)$$

When using a base 2 logarithm, the information is measured in bits or binary units.

MEASURES OF SHAPE

Measures of skewness $\sqrt{\beta_1}(X)$ and excess kurtosis $\beta_2(X)-3$ based on standardized central moments with the original Karl Pearson notation:

$$\sqrt{\beta_1(X)} = \frac{E\left[(X - E(X))^3\right]}{\sqrt{E^3\left[(X - E(X))^2\right]}} = \left(e^{\sigma^2} + 2\right)\sqrt{e^{\sigma^2} - 1} = \left(e^{\sigma^2} + 2\right)CV(X)$$

$$\beta_2(X) - 3 = \frac{E\left[(X - E(X))^4\right]}{E^2\left[(X - E(X))^2\right]} = e^{4\sigma^2} + 2e^{3\sigma^2} + 3e^{2\sigma^2} - 6 = \mu_g^4(X) + 2\mu_g^3(X) + 3\mu_g^2(X) - 6$$

ESTIMATORS OF μ AND σ^2 FROM THE LOGNORMAL DISTRIBUTION

Consider a variable X with lognormal distribution of unknown parameters μ and σ^2. A sample of size *n* is randomly drawn. The estimators by the method of moments are based on the arithmetic mean and variance of the *n* untransformed sample data (Stuart and Ord, 2010).

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n};\ s_n^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

$$\hat{\mu} = \ln\left(\frac{E(X)}{\sqrt{1 + \frac{Var(X)}{E^2(X)}}}\right) = \ln\left(\frac{\bar{x}}{\sqrt{1 + \frac{s_n^2}{\bar{x}^2}}}\right) = \ln\left(\frac{\bar{x}^2}{\sqrt{\bar{x}^2 + s_n^2}}\right)$$

$$\hat{\sigma}^2 = \ln\left(1 + \frac{Var(X)}{E^2(X)}\right) = \ln\left(1 + \frac{s_n^2}{\bar{x}^2}\right) = \ln\left(\frac{\bar{x}^2 + s_n^2}{\bar{x}^2}\right)$$

Maximum likelihood estimators use the mean and variance of the data transformed logarithmically with natural basis.

$$\hat{\mu} = \overline{\ln(x_i)} = \frac{\sum_{i=1}^{n} \ln(x_i)}{n}$$

$$\hat{\sigma}^2 = s_{ln(x)}^2 = \frac{\sum_{i=1}^{n}\left(\ln(x_i) - \overline{\ln(x)}\right)^2}{n}$$

The maximum likelihood estimator of μ is unbiased, but not that of σ^2. A bias correction for the latter consists of dividing by *n* - 1 instead of *n*.

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}\left(\ln(x_i) - \overline{\ln(x)}\right)^2}{n - 1}$$

Asymptotic properties that maximum likelihood estimators possess include consistency, normality, and efficiency. These properties allow obtaining the asymptotic standard error (*ASE*) from Fisher's information for *n* data (or the inverse of Cramer-Rao lower bound) when the estimator is unbiased, as is the case with the μ estimator. It also allows for defining a Wald-type confidence interval (Lauritzen et al., 2019).

$$I(\mu,\sigma^2) = n\begin{pmatrix} -E\left[\frac{\delta}{\delta\mu\delta\mu}\ln f_X(x|\mu,\sigma^2)\right] & -E\left[\frac{\delta}{\delta\mu\delta\sigma^2}\ln f_X(x|\mu,\sigma^2)\right] \\ -E\left[\frac{\delta}{\delta\sigma^2\delta\mu}\ln f_X(x|\mu,\sigma^2)\right] & -E\left[\frac{\delta}{\delta\sigma^2\delta\sigma^2}\ln f_X(x|\mu,\sigma^2)\right] \end{pmatrix} = \begin{pmatrix} n/\sigma^2 & 0 \\ 0 & 2n/\sigma^2 \end{pmatrix}$$

$$\sigma_{\hat{\mu}}^2 = (1/I(\mu))^{-1} = \sigma^2/n;\ \sigma_{\hat{\mu}} = \sqrt{\sigma^2/n}$$

$$P\left(\hat{\mu} - z_{1-\frac{\alpha}{2}}\sigma_{\hat{\mu}}^2 \leq \mu \leq \hat{\mu} + z_{1-\frac{\alpha}{2}}\sigma_{\hat{\mu}}^2\right) = P\left(\hat{\mu} - z_{1-\frac{\alpha}{2}}\sqrt{\hat{\sigma}^2/n} \leq \mu \leq \hat{\mu} + z_{1-\frac{\alpha}{2}}\sqrt{\hat{\sigma}^2/n}\right) = 1 - \alpha$$

$z_{1-\alpha/2}$ = $\Phi^{-1}(1-\alpha/2)$ = quantile of order 1 – α/2 from standard normal distribution.

$$\hat{\sigma}_{\hat{\mu}}^2 = (1/I(\mu))^{-1} = \hat{\sigma}^2/n = s_{\ln(x)}^2/n;\ ASE = \hat{\sigma}_{\hat{\mu}} = \sqrt{\hat{\sigma}^2/n} = \sqrt{s_{\ln(x)}^2/n}$$

$$P\left(\overline{\ln(x)} - z_{1-\frac{\alpha}{2}}\sqrt{s_{\ln(x)}^2/n} \leq \mu \leq \overline{\ln(x)} + z_{1-\frac{\alpha}{2}}\sqrt{s_{\ln(x)}^2/n}\right) = 1 - \alpha$$

PROPERTIES OF THE LOG-NORMAL DISTRIBUTION

The arithmetic mean is the ratio of the square of the geometric mean to the harmonic mean. By rearranging this equation, we find that the harmonic mean is the ratio of the square of the geometric mean to the arithmetic mean. Similarly, the geometric mean is the square root of the product of the arithmetic mean and the harmonic mean.

$$E(X) = \frac{[G(X)]^2}{H(X)};\ H(X) = \frac{[G(X)]^2}{M(X)};\ G(X) = \sqrt{E(X)H(X)}$$

If the random variable X ~ *Lognormal* (μ, σ^2) and the constant *a* > 0, then *a* × X ~ *Lognormal* (μ + ln(*a*), σ^2). If the random variable X ~ *Lognormal* (μ, σ2), then 1/X ~ *Lognormal* (–μ, σ^2). If the random variable X ~ *Lognormal* (μ, σ^2) and the constant *a* ≠ 0, then X^a ~ *Lognormal* (*a* × μ, a^2 × σ^2). If X_1 ~ *Lognormal* (μ_1, σ_1^2) and X_2 ~ *Lognormal* (μ_2, σ_2^2) and both random variables are independent, then the product X_1 × X_2 ~ *Lognormal* (μ_1 + μ_2, σ_1^2 + σ_2^2) and the quotient X_1 / X_2 ~ *Lognormal* (μ_1 − μ_2, σ_1^2 + σ_2^2). If X_1 ~ *Lognormal* (μ_1, σ_1^2), X_2 ~ *Lognormal* (μ_2, σ_2^2), ..., X_n ~ *Lognormal* (μ_n, σ_n^2) and the *n* variables are mutually independent, then the product X_1 × X_2 × … × X_n ~ *Lognormal* (μ_1 + μ_2 + … + μ_n, σ_1^2 + σ_2^2 + … + σ_n^2).

Let X_1, X_2, …, X_n be *n* random, independent, and identically distributed variables with a finite mean and variance. Then, their geometric mean follows a lognormal distribution, where the location parameter μ is the arithmetic mean or mathematical expectation of the log-transformed variables with the natural base, and its squared scale parameter σ^2 is the quotient between the variance of the log-transformed variables with the natural base and the number of variables. This property constitutes the so-called multiplicative central limit theorem (Galton, 1879). Another expression of the theorem is Gibrat's law (1931), which holds true when a natural growth process results from the accumulation of small multiplicative changes that, when transformed to a logarithmic scale, become additive increments. This concept finds applications in economics (Balthrop, 2021; Ahundjanov and Toda, 2020), demography (Ciccone, 2021), and environment (Ahundjanov and Akhundjanov, 2019). Consequently, the product of *n* independent random variables with finite means and variances follows a lognormal distribution when their sum follows a normal distribution (Haines et al., 2020; Zijian, 2020).

$$X = \{X_1, X_2, \dots, X_n\}$$

$$G(X) = \mu_g(X) = \sqrt[n]{\prod_{i=1}^{n} X_i} = e^{\sum_{i=1}^{n} \ln(X_i)/n}$$

$$G(X) \sim Lognormal\left(\mu = E[\ln(X)], \sigma^2 = \frac{Var[\ln(X)]}{n}\right)$$

$$E[\ln(X)] = \frac{\sum_{i=1}^{n} \ln(X_i)}{n}$$

$$Var[\ln(X)] = E[(\ln(x) - E[\ln(x)])^2] = \frac{\sum_{i=1}^{n} (\ln(X_i) - E[\ln(X)])^2}{n}$$

In a lognormal distribution, 68.3% of the data lies in the interval between the quotient of the geometric mean and the geometric standard deviation (lower limit) and the product of the geometric mean and the geometric standard deviation (upper limit). The 95.4% of the data falls within the range bounded by the quotient of the geometric mean and the square of the geometric standard deviation (lower bound) and the product of the geometric mean and the square of the geometric standard deviation (upper bound). The 99.7% of the data is in the interval between the quotient of the geometric mean and the cube of the geometric standard deviation (lower limit) and the product of the geometric mean and the cube of the geometric standard deviation (upper limit).

As seen previously, the geometric mean can be calculated as the exponential function with base *e* and an exponent equal to the arithmetic mean of the logarithmically transformed values with a natural base. The geometric standard deviation is obtained by applying the exponential function with base *e* to an exponent equal to the standard deviation of the logarithmically transformed values with a natural base. It is important to note that the geometric variance is computed using the exponential function with base *e* and an exponent equal to the variance of the logarithmically transformed values with a natural base. This results in a value different from the square of the geometric standard deviation.

Mean and standard deviation of log-transformed data at the population level (μ and σ):

$$\mu_{\ln(x)} = E[\ln(x)]$$
$$\sigma_{\ln(x)} = E[(\ln(x) - E[\ln(x)])^2]$$

Geometric mean and standard deviation of X data at population level (μ_g and σ_g):

$$\mu_g(X) = e^{\mu_{\ln(x)}}$$
$$\sigma_g(X) = e^{\sigma_{\ln(x)}}$$

Dispersion intervals:

$$P(\ln(x) \in [\mu_{\ln(x)} - \sigma_{\ln(x)}, \mu_{\ln(x)} + \sigma_{\ln(x)}])$$

$$= P(x \in [\mu_g(X)/\sigma_g(X), \mu_g(X) \times \sigma_g(X)]) = P(x \in [e^{\mu_{\ln(x)} - \sigma_{\ln(x)}}, e^{\mu_{\ln(x)} + \sigma_{\ln(x)}}]) = 0.683$$

$$P(\ln(x) \in [\mu_{\ln(x)} - 2\sigma_{\ln(x)}, \mu_{\ln(x)} + 2\sigma_{\ln(x)}])$$

$$= P\left(x \in \left[\mu_g(X)/[\sigma_g(X)]^2, \mu_g(X) \times [\sigma_g(X)]^2\right]\right) = P\left(x \in \left[e^{\mu_{\ln(x)} - \sigma_{\ln(x)}^2}, e^{\mu_{\ln(x)} + \sigma_{\ln(x)}^2}\right]\right) = 0.954$$

$$P(\ln(x) \in [\mu_{\ln(x)} - 3\sigma_{\ln(x)}, \mu_{\ln(x)} + 3\sigma_{\ln(x)}])$$

$$= P\left(x \in \left[\mu_g(X)/[\sigma_g(X)]^2, \mu_g(X) \times [\sigma_g(X)]^2\right]\right) = P\left(x \in \left[e^{\mu_{\ln(x)} - \sigma_{\ln(x)}^3}, e^{\mu_{\ln(x)} + \sigma_{\ln(x)}^3}\right]\right) = 0.997$$

Considering the dispersion intervals of the data for a variable X in a lognormal distribution, confidence intervals can be defined for the geometric mean of this distribution. Let a random sample of *n* data points from a continuous variable X with a lognormal distribution characterized by unknown parameters μ and σ^2.

Unknown population geometric mean:

$$X \sim Lognormal(\mu, \sigma^2)$$
$$\mu_g(X) = Mdn(X) = e^{\mu} = e^{\mu_{\ln(X)}}$$

Sample geometric mean:

$$\hat{\mu}_g(X) = e^{\hat{\mu}_{\ln(x)}} = e^{m_{\ln(x)}} = e^{\overline{\ln(x)}} = e^{\sum_{i=1}^{n} \ln(x_i)/n}$$

Standard error of the sample geometric mean:

$$EEG = \left(\hat{\sigma}_g(X)\right)^{1/\sqrt{n}} = \left(e^{\hat{\sigma}_{\ln(x)}}\right)^{1/\sqrt{n}} = e^{\frac{\hat{\sigma}_{\ln(x)}}{\sqrt{n}}} = e^{\frac{s_{\ln(x)}}{\sqrt{n}}} = e^{\frac{\sqrt{\sum_{i=1}^{n} (\ln(x_i) - \overline{\ln(x)})^2}}{n}}$$

Confidence interval at 1 – α for the geometric mean:

$$P\left(\hat{\mu}_g(X)/\left(\hat{\sigma}_g(X)\right)^{\frac{{}_{1-\alpha/2}t_{n-1}}{\sqrt{n}}} \leq \mu_g(X) \leq \hat{\mu}_g(X)\left(\hat{\sigma}_g(X)\right)^{\frac{{}_{1-\alpha/2}t_{n-1}}{\sqrt{n}}}\right)$$

$$= P\left(e^{\hat{\mu}_{\ln(x)}}/\left(e^{\hat{\sigma}_{\ln(x)}}\right)^{\frac{{}_{1-\alpha/2}t_{n-1}}{\sqrt{n}}} \leq \mu_g(X) \leq e^{\hat{\mu}_{\ln(x)}}\left(e^{\hat{\sigma}_{\ln(x)}}\right)^{\frac{{}_{1-\alpha/2}t_{n-1}}{\sqrt{n}}}\right)$$

$$= P\left(e^{\hat{\mu}_{\ln(x)}}/e^{{}_{1-\alpha/2}t_{n-1}\frac{\hat{\sigma}_{\ln(x)}}{\sqrt{n}}} \leq \mu_g(X) \leq e^{\hat{\mu}_{\ln(x)}} e^{{}_{1-\alpha/2}t_{n-1}\frac{\hat{\sigma}_{\ln(x)}}{\sqrt{n}}}\right)$$

$$= P\left(e^{\hat{\mu}_{\ln(x)} - {}_{1-\alpha/2}t_{n-1}\frac{\hat{\sigma}_{\ln(x)}}{\sqrt{n}}} \leq e^{\mu_{\ln(X)}} \leq e^{\hat{\mu}_{\ln(x)} + {}_{1-\alpha/2}t_{n-1}\frac{\hat{\sigma}_{\ln(x)}}{\sqrt{n}}}\right) = 1 - \alpha$$

where ${}_{1-\alpha/2}t_{n-1}$ = quantile of order 1 – α/2 of Student's t-distribution with *n* – 1 degrees of freedom.

To generate a random sample from a continuous variable with a lognormal distribution with parameters α and σ^2, you can obtain a random sample from a variable with a standard normal distribution and apply an exponential transformation.

$$z \in Z \sim N(0,1)$$

$$x \in X = e^{\mu+\sigma z} \sim Lognormal(\mu, \sigma^2)$$

To calculate the probability values of a variable with a lognormal distribution with parameters μ and σ^2, the process involves first transforming the values using the natural logarithm, then standardizing the log-transformed values, and finally using the cumulative distribution function $F_X(x)$ of the standard normal distribution *N* (0, 1).

$$X \sim Lognormal(\mu, \sigma^2)$$

$$\mu = E[\ln(X)] = \ln[G(X)]$$

$$G(X) = Mdn(X) = e^{\mu} = e^{\mu_{\ln(X)}}$$

$$\sigma = SD(\ln(X)) = \sqrt{E\left[\left(\ln(X) - E(\ln(X))\right)^2\right]} = \ln[GSD(X)]$$

$$GSD(X) = e^{\sigma} = e^{\sigma_{\ln(X)}}$$

$$Z = \frac{\ln(x) - \mu}{\sigma} = \frac{\ln(x) - \ln[G(X)]}{\ln[GSD(X)]} = \frac{\ln(x) - \ln[e^{\mu_{\ln(X)}}]}{\ln[e^{\sigma_{\ln(X)}}]} = \frac{\ln(x) - \mu_{\ln(X)}}{\sigma_{\ln(X)}} \sim N(0,1)$$

$$P(X \leq x) = P\left(Z \leq \frac{\ln(x) - \mu}{\sigma}\right)$$

$$P(X \geq x) = 1 - P\left(Z \leq \frac{\ln(x) - \mu}{\sigma}\right)$$

$$P(x_1 \leq X \leq x_2) = P\left(Z \leq \frac{\ln(x_2) - \mu}{\sigma}\right) - P\left(Z \leq \frac{\ln(x_1) - \mu}{\sigma}\right)$$

The following expression is referred to as the geometrically standardized score (Finlay and Darquenne, 2020):

$$z_g(x) = \frac{\ln(x) - \mu}{\sigma} = \frac{\ln(x) - \ln[\mu_g(X)]}{\ln[\sigma_g(X)]} = \frac{\ln\left(\frac{x}{\mu_g(X)}\right)}{\ln[\sigma_g(X)]} = \log_{\sigma_g(x)}\left(\frac{x}{\mu_g(X)}\right)$$

USEFULNESS OF THE LOGNORMAL DISTRIBUTION IN RESEARCH AND THEORETICAL FORMULATION

The lognormal distribution is highly useful in research and theoretical formulations across various fields, including the social and health sciences, as well as in other scientific disciplines. The following points highlight its applications and importance in these fields.

Social sciences

Income and wealth distribution

Income distribution: In economics, the distribution of income among individuals or households often follows a lognormal distribution. This is because income can be thought of as the product of many independent factors (education, experience, etc.), which multiplicatively combine to produce the overall income, a scenario well-modeled by the lognormal distribution (Özkan et al., 2020).

Wealth distribution: Similarly, the distribution of wealth is often modeled as lognormal, reflecting the multiplicative effects of savings, investment returns, inheritance, and other factors (Akhundjanov and Toda, 2020; Eguia and Xefteris, 2022).

Behavioral and social phenomena

Sexual behavior: The lognormal distribution has been used for modelling number of sexual partners (Major et al., 2021), age at first sex among youth (Materu et al., 2023), and age at first forced sexual act among women (Kawuki et al., 2021) among other topics of sexual behavior.

Internet use and online behavior: The frequency of internet usage, the distribution of followers on social media platforms (Aggrawal et al., 2021), and other online behaviors often exhibit lognormal characteristics due to multiplicative growth effects (Alasmar et al., 2021; Pal et al., 2021).

City sizes: The sizes of cities in a country or region often follow a lognormal distribution, as they result from multiplicative growth processes, as birth rates and migration patterns (González-Val, 2021; Verbavatz and Barthelemy, 2020).

Health sciences

Biomedical data

Physiological measurements: Variables like blood pressure (Chalkias and Xenos, 2022), prolactin secretion

(Gayathri et al., 2022), and sizes of tumors (Chan et al., 2021; Wang et al., 2021) often follow a lognormal distribution. These are influenced by a multiplicative combination of genetic, environmental, and lifestyle factors (Fossion et al., 2020).

Survival and failure times: The lognormal distribution is useful in modeling survival times and life expectancy, especially in situations where the rate of aging or failure is proportional to the current age or size (Maccone, 2020; Olosunde and Ejiofor, 2021; Waymyers and Chakraborty, 2024).

Reaction times: In the field of psychology, the lognormal distribution has been applied to modeling responses and response times in tests (Ranger et al., 2020; Sinharay and van Rijn, 2020).

Epidemiology

Disease spread: The spread of diseases can sometimes be better modeled by a lognormal distribution, particularly when considering the time until infection or recovery, due to the multiplicative nature of biological and environmental interactions. Recently, it has been applied to the COVID-19 (SARS-CoV-2) epidemic, with the studies for forecasting the spread of Covid-19 (Lawrence et al., 2023), incubation period of COVID-19 (McAloon et al., 2020), and predicting COVID-19 deaths (Valvo, 2020).

Theoretical formulation

Multiplicative processes: Many social and health phenomena result from multiplicative processes. The lognormal distribution is theoretically appropriate when the underlying processes can be modeled as products of random variables (Chen and Korsunsky, 2021).

Modeling variability and risk: In both social and health sciences, understanding variability and risk is crucial. The lognormal distribution allows researchers to model skewed data with a heavy tail, which is common in these fields. This is important for risk assessment and decision-making (Guo and Li, 2021).

Advantages of the lognormal distribution

Flexibility and realism: The lognormal distribution is flexible and can realistically model a wide range of phenomena characterized by positive skewness and a long tail (Goldenholz and Westover, 2023).

Parameter interpretation: Parameters of the lognormal distribution (mean and variance in the logarithmic scale) can be interpreted in terms of multiplicative factors, making them intuitive for modeling and hypothesis testing in social and health contexts (Andersson, 2021; Elassaiss-Schaap and Duisters, 2020).

Data transformation: Many statistical methods assume normality of data. Transforming lognormal data (by taking the natural logarithm) can simplify analysis and make it amenable to standard techniques (Choi et al., 2022).

Practical applications

Policy and planning: In public health, understanding the distribution of health outcomes can inform interventions and resource allocation. In economics, knowledge of income distribution aids in policy formulation for taxation and welfare (Aslam, 2024; Balci and Kumral, 2024; Beykaei et al., 2020; Phelps, 2023).

Risk management: In both health and social sciences, the lognormal distribution helps in risk management by modeling extreme events and long-tail risks, such as catastrophic health expenses or economic downturns (Guo and Li, 2021; Jokhadze and Schmidt, 2020).

Thus, the lognormal distribution is a powerful tool in social and health sciences due to its ability to model complex, multiplicative processes and its applicability to a wide range of empirical data. Its use enhances the accuracy and interpretability of research findings, informing theory development and practical applications in these fields.

CALCULATION EXAMPLES WITH THE LOGNORMAL DISTRIBUTION

Both examples were generated specifically for this article. In the second example, first a random sample of 30 data was drawn from a continuous uniform distribution with the Excel random number generator. From this, a standard normal distribution was obtained using the probit function, and through an exponential transformation, a random sample was created. To make the examples more meaningful, they are given content related to sexual behavior and presented as if they were empirical data.

Example 1: With the known probability distribution and its specified parameters

In a population of young Mexican women with at least seven years of marriage, we inquired about the average number of sexual relations per month in the last year, encompassing both marital and concurrent partners. It was observed that the variable distribution follows a lognormal distribution with parameters $\mu = 0.7$ and $\sigma^2 =$

1.28. What are the probabilities that a woman has an average number of sexual relations per month of three or more, one or fewer, and between one and two and a half? Calculate the arithmetic, geometric, and harmonic means, along with the median and mode, as measures of central tendency for this distribution. Compute the variance, standard deviation, and coefficient of variation, both arithmetic and geometric, as well as the Shannon's entropy as measures of variation. Compute measures of skewness and excess kurtosis based on standardized central moments. Determine the dispersion intervals for approximately 68.3, 95.4, and 99.7% of the data. Finally, plot the cumulative density and distribution function.

X = average number of sexual relations per month in the last year in young Mexican women with at least seven years of marriage, including concurrent partners.

X ~ *Lognormal* (μ = 0.7, σ^2 = 1.28)

What is the probability that a woman has an average of three or more sexual relations per month?

μ = 0.7, σ^2 = 1.28, σ = √1.28 = 1.13

$$P(X \geq x = 3) = P\left(Z \geq \frac{\ln(x) - \mu}{\sigma}\right) = 1 - P\left(Z < \frac{\ln(x) - \mu}{\sigma}\right) = 1 - P\left(Z < \frac{\ln(3) - 0.7}{1.13}\right)$$
$$= 1 - P(Z < 0.35) = 1 - 0.64 = 0.36$$

What is the probability that a woman has an average number of sexual relations per month of one or less?

$$P(X \leq 1) = P\left(Z \leq \frac{\ln(x) - \mu}{\sigma}\right) = P\left(Z \leq \frac{\ln(1) - 0.7}{1.13}\right) = P(Z \leq -0.62) = 0.27$$

What is the probability that a woman has an average of one to two and a half sexual relations per month?

$$P(1 \leq X \leq 2.5) = P(X \leq 2.5) - P(X \leq 1) = P\left(Z \leq \frac{\ln(2.5) - 0.7}{1.13}\right) - P\left(Z \leq \frac{\ln(1) - 0.7}{1.13}\right)$$
$$= P(Z \leq 0.19) - P(Z \leq -0.62) = 0.58 - 0.27 = 0.31$$

Calculation of the arithmetic mean, μ(X), geometric mean, μ$_g$(X), harmonic mean, μ$_h$(X), median, *Mdn*(X), and mode, *Mo*(X) as measures of central tendency:

$$E(X) = \mu(X) = e^{\mu + \sigma^2/2} = e^{0.7 + \frac{1.28}{2}} = 3.82$$

$$E(X) = \mu(X) = \frac{[G(X)]^2}{H(X)} = \frac{[\mu_g(X)]^2}{\mu_h(X)} = \frac{2.01^2}{1.06} = 3.82$$

$$G(X) = \mu_g(X) = e^{\mu} = e^{0.7} = 2.01$$

$$G(X) = \mu_g(X) = \sqrt{\mu(X) \times \mu_h(X)} = \sqrt{3.82 \times 1.06} = 2.01$$

$$H(X) = \mu_h(X) = e^{\mu - \sigma^2/2} = e^{0.7 - \frac{1.28}{2}} = 1.06$$

$$H(X) = \mu_h(X) = \frac{[G(X)]^2}{E(X)} = \frac{[\mu_g(X)]^2}{\mu(X)} = \frac{2.01^2}{3.82} = 1.06$$

$$Mdn(X) = Q_X(p = 1/2) = e^{\mu} = e^{0.7} = 2.01$$

$$P(X \leq 2.0138) = P\left(Z \leq \frac{\ln(x) - \mu}{\sigma}\right) = P\left(Z \leq \frac{\ln(2.01) - 0.7}{1.13}\right) = P(Z \leq 0) = 0.5$$

$$Mo(X) = e^{\mu - \sigma^2} = e^{0.7 - 1.28} = 0.56$$

Calculation of arithmetic variance (σ^2), standard deviation (σ), and coefficient of variation (*CV*), as well as geometric variance (σ$_g$^{2}), standard deviation (σ$_g$), and coefficient of variation (*GCV*), along with Shannon's entropy (H) as measures of variation:

$$Var(X) = \sigma^2(X) = e^{2\mu + \sigma^2}\left(e^{\sigma^2} - 1\right) = e^{2 \times 0.7 + 1.28}\left(e^{1.28} - 1\right) = 37.87$$

$$SD(X) = \sigma(X) = e^{\mu + \frac{\sigma^2}{2}}\sqrt{e^{\sigma^2} - 1} = e^{0.7 + \frac{1.28}{2}}\sqrt{e^{1.28} - 1} = 6.15$$

$$SD(X) = E(X) \times CV(X) = 3.82 \times 1.61 = 6.15$$

$$SD(X) = \sqrt{Var(X)} = \sqrt{37.87} = 6.15$$

$$CV(X) = \frac{SD(X)}{|E(X)|} = \sqrt{e^{\sigma^2} - 1} = \sqrt{e^{1.28} - 1} = 1.61$$

$$GV(X) = \sigma_g^2(X) = E(\ln(X) - E[\ln(X)]) = e^{\sigma^2_{\ln(X)}} = e^{\sigma^2} = e^{1.28} = 3.60$$

$$GSD(X) = \sigma_g(X) = \sqrt{E(\ln(X) - E[\ln(X)])} = e^{\sigma_{\ln(X)}} = e^{\sigma} = e^{1.13} = 3.10$$

$$GCV = \sigma_g(X) - 1 = e^{\sigma} - 1 = e^{1.13} - 1 = 2.10$$

$$\mathrm{H}(X) = E(-\log_2[f_X(x)]) = \log_2\left(\sigma e^{\mu + \frac{1}{2}}\sqrt{2\pi}\right) = \log_2\left(1.13 \times e^{0.7 + \frac{1}{2}} \times \sqrt{2\pi}\right) = \log_2(9.42)$$
$$= \frac{\ln(9.42)}{\ln(2)} = 3.24\ bits$$

Calculation of measures of skewness and excess kurtosis based on standardized central moments as measures of the shape of the distribution.

$$\sqrt{\beta_1(X)} = \left(e^{\sigma^2} + 2\right) \times \sqrt{e^{\sigma^2} - 1} = \left(e^{1.28} + 2\right) \times \sqrt{e^{1.28} - 1} = 9.02$$

$$\beta_2(X) - 3 = e^{4\sigma^2} + 2e^{3\sigma^2} + 3e^{2\sigma^2} - 6 = e^{4 \mathrm{X} 1.28} + 2e^{3 \mathrm{X} 1.28} + 3e^{2 \mathrm{X} 1.28} - 6 = 293.19$$

Testing of the previously defined dispersion intervals for approximately 68.3, 95.4 and 99.7% of the data:

$$P\left(x \in [\mu_g(X)/\sigma_g(X), \mu_g(X) \times \sigma_g(X)]\right) = 0.683$$

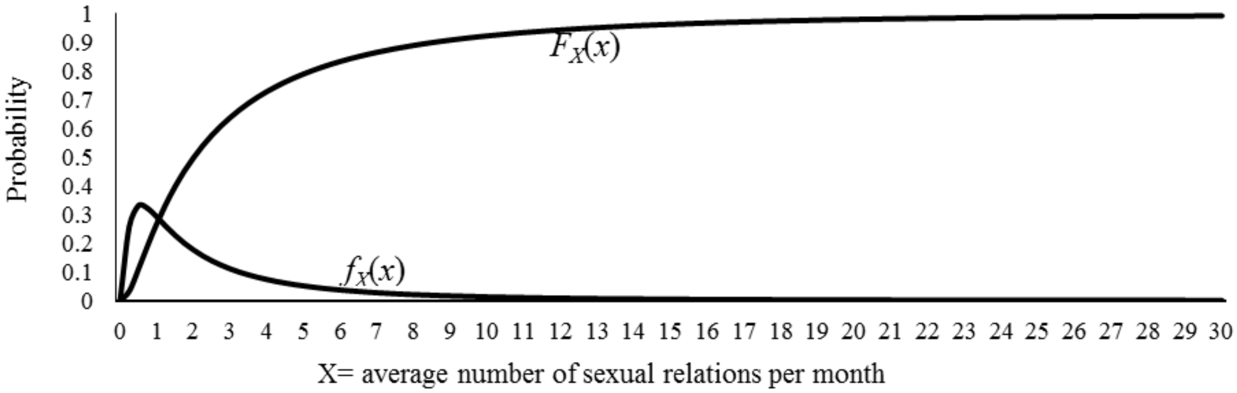

Figure 3. Plot of the probability density function $f_X(x)$ and cumulative distribution function $F_X(x)$ of X ~ Lognormal (μ = 0.7, σ^2 = 1.28).
Source: Author's elaboration.

$$P\left(x \in \left[\frac{2.01}{3.10}, 2.01 \times 3.10\right]\right) = P(x \in [0.65, 6.24])$$

$$= P\left(z = \frac{\ln(x) - \mu_{\ln(x)}}{\sigma_{\ln(x)}} \in \left[\frac{\ln(0.65) - 0.7}{1.13}, \frac{\ln(6.24) - 0.7}{1.13}\right]\right) = P(Z \in [-1,1]) = 0.683$$

$$P\left(x \in [\mu_g(X)/\sigma_g^2(X), \mu_g(X) \times \sigma_g^2(X)]\right) = 0.954$$

$$P\left(x \in \left[\frac{2.01}{3.10^2}, 2.01 \times 3.10^2\right]\right) = P(x \in [0.21, 19.35])$$

$$= P\left(z = \frac{\ln(x) - \mu_{\ln(x)}}{\sigma_{\ln(x)}} \in \left[\frac{\ln(0.21) - 0.7}{1.13}, \frac{\ln(19.35) - 0.7}{1.13}\right]\right) = P(Z \in [-2,2]) = 0.954$$

$$P\left(x \in [\mu_g(X)/\sigma_g^3(X), \mu_g(X) \times \sigma_g^3(X)]\right) = 0.997$$

$$P\left(x \in \left[\frac{2.01}{3.10^3}, 2.01 \times 3.10^3\right]\right) = P(x \in [0.07, 59.99])$$

$$= P\left(z = \frac{\ln(x) - \mu_{\ln(x)}}{\sigma_{\ln(x)}} \in \left[\frac{\ln(0.07) - 0.7}{1.13}, \frac{\ln(59.99) - 0.7}{1.13}\right]\right) = P(Z \in [-3,3]) = 0.997$$

Figure 3 depicts the density and cumulative distribution function of the variable representing the average number of sexual relations per month, which follows a lognormal distribution with a location parameter μ = 0.7 and squared scale σ² = 1.28.

Example 2: With the hypothesized probability distribution and its estimated parameters

Be a random sample of 30 participants from the population of young Mexican women with at least seven years of marriage. We measured the average number of sexual relations per month over the past year, including both marital and concurrent partners, through a questionnaire (Table 1). This positive continuous quantitative variable follows a lognormal distribution. Make the point estimate of its parameters μ and σ² (on a logarithmic scale), estimate the 95% confidence interval for its location parameter μ, and determine the median number of sexual intercourses along with its geometric variance and standard deviation. Estimate the geometric mean with a 95% confidence interval. Calculate the probability of having an average of at least 8 sexual relations per month and the density of having an average of 8 sexual relations per month. Determine the average corresponding to a cumulative probability of 90%. Finally, plot the density and cumulative distribution function using the 30 sample data and the parameter estimates obtained through the maximum likelihood method.

The x_i data in Table 1 were generated through simulation from a standard normal distribution using the Excel 2021 random number generator. The initial sequence of normal data was adjusted to have a mean and variance exactly equal to 0 and 1, respectively. To achieve this, the sample mean was subtracted, and the difference was divided by the uncorrected standard deviation to correct for bias. Subsequently, these normal data, denoted as z_i with a mean of 0 and variance of 1, underwent transformation: $x_i = e^{0.7+\sqrt{1.28} \times z_i}$. This transformation resulted in the creation of a random sample with a lognormal distribution with parameters μ = 0.7 and σ² = 1.28.

Estimation of parameters by the method of moments and calculation of the geometric mean and variance and standard deviation:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{108.77}{30} = 3.63$$

$$s_n^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = \frac{551.20}{30} = 18.37$$

Table 1. Average number of sexual relations per month for the 30 participants, their logarithmic transformation, squared differential log-transformed scores, probability density function, cumulative distribution function, and theoretical quantiles.

i	(i)	$x_{(i)}$	$[x_{(i)}-E(X)]^2$	$y_{(i)}=\ln(x_{(i)})$	$[y_{(i)}-E(y_i)]^2$	$f_X(x_{(i)})$	$F_X(x_{(i)})$	$p_{(i)}$	$Q_X(p_{(i)})$
29	1	0.1109	12.3538	-2.1989	8.4036	0.1193	0.0052	0.0220	0.2061
17	2	0.4979	9.7834	-0.6974	1.9527	0.3303	0.1084	0.0549	0.3300
4	3	0.5444	9.4945	-0.6081	1.7111	0.3320	0.1238	0.0879	0.4354
22	4	0.5555	9.4264	-0.5879	1.6588	0.3321	0.1275	0.1209	0.5356
16	5	0.5646	9.3703	-0.5716	1.6169	0.3321	0.1305	0.1538	0.6350
19	6	0.6131	9.0756	-0.4892	1.4141	0.3310	0.1466	0.1868	0.7360
12	7	0.7720	8.1436	-0.2588	0.9192	0.3190	0.1984	0.2198	0.8399
8	8	0.9779	7.0107	-0.0223	0.5217	0.2941	0.2616	0.2527	0.9481
7	9	1.2169	5.8025	0.1963	0.2537	0.2624	0.3281	0.2857	1.0615
10	10	1.2593	5.5998	0.2306	0.2204	0.2569	0.3391	0.3187	1.1814
24	11	1.3489	5.1841	0.2993	0.1606	0.2455	0.3616	0.3516	1.3088
28	12	1.3822	5.0334	0.3237	0.1416	0.2414	0.3697	0.3846	1.4450
20	13	1.4422	4.7677	0.3662	0.1114	0.2341	0.3840	0.4176	1.5913
6	14	1.5129	4.4642	0.4140	0.0818	0.2258	0.4002	0.4505	1.7496
11	15	1.5713	4.2207	0.4519	0.0616	0.2191	0.4132	0.4835	1.9218
27	16	1.5875	4.1544	0.4621	0.0566	0.2173	0.4167	0.5165	2.1102
5	17	2.2109	2.0017	0.7934	0.0087	0.1589	0.5329	0.5495	2.3178
15	18	2.3390	1.6557	0.8497	0.0224	0.1494	0.5526	0.5824	2.5483
14	19	2.4982	1.2714	0.9156	0.0465	0.1386	0.5755	0.6154	2.8065
18	20	3.0408	0.3421	1.1121	0.1698	0.1085	0.6422	0.6484	3.0985
1	21	3.8358	0.0441	1.3444	0.4152	0.0782	0.7155	0.6813	3.4326
13	22	5.0605	2.0585	1.6215	0.8491	0.0500	0.7923	0.7143	3.8202
25	23	6.1116	6.1795	1.8102	1.2325	0.0357	0.8368	0.7473	4.2774
26	24	6.4789	8.1404	1.8685	1.3655	0.0319	0.8492	0.7802	4.8282
21	25	6.4860	8.1815	1.8697	1.3681	0.0319	0.8494	0.8132	5.5101
2	26	6.7360	9.6741	1.9075	1.4580	0.0296	0.8571	0.8462	6.3858
9	27	7.8135	17.5371	2.0558	1.8383	0.0220	0.8846	0.8791	7.5715
3	28	8.2864	21.7222	2.1146	2.0011	0.0195	0.8944	0.9121	9.3144
30	29	10.7466	50.7069	2.3746	2.8042	0.0110	0.9306	0.9451	12.2890
23	30	21.1700	307.8021	3.0526	5.5347	0.0019	0.9812	0.9780	19.6712
Σ		108.7715	551.2026	21	38.4				

i = random order, $x_{(i)}$ = number of sexual relations or scores in X with data sorted in ascending order (i = 1, 2, ..., 30), $[x_{(i)} - E(X)]^2$ = square of the differential scores (with respect to the arithmetic mean) of X with data sorted in ascending order, $y_{(i)} = \ln(x_{(i)})$ = natural-based log transformation of the scores of X with the data sorted in ascending order, $[y_{(i)} - E(y_i)]^2 = [\ln(x_{(i)} - E(\ln(X))]^2$ = square of the differential scores (with respect to the arithmetic mean) of the logarithms of the scores of X, $f_X(x_{(i)})$ = probability density function and $F_X(x_{(i)})$ = cumulative distribution function calculated using the sample data $x_{(i)}$ under the lognormal distribution with parameters $\mu = 0.7$ and $\sigma^2 = 1.28$, $p_{(i)} = [(i) - 1/3] / (30 + 1/3)$ = theoretical quantile order, $Q_X(p_{(i)})$ = theoretical quantiles under the lognormal distribution with parameters $\mu = 0.7$ and $\sigma^2 = 1.28$, and Σ = sum per column.
Source: Author's elaboration.

$$\hat{\mu} = \ln\left(\frac{\bar{x}^2}{\sqrt{\bar{x}^2 + s_n^2}}\right) = \ln\left(\frac{3.63^2}{\sqrt{3.63^2 + 18.37}}\right) = \ln(2.34) = 0.85$$

$$\widehat{Mdn}(X) = \hat{G}(X) = \widehat{\mu_g}(X) = e^{\hat{\mu}} = e^{0.85} = 2.34$$

$$\hat{\sigma}^2 = \ln\left(1 + \frac{s_n^2}{\bar{x}^2}\right) = \ln\left(1 + \frac{18.37}{3.63^2}\right) = \ln(2.40) = 0.87$$

$$\hat{\sigma}_g^2(X) = e^{\hat{\sigma}^2} = e^{0.87} = 2.40$$

$$\hat{\sigma} = \sqrt{\ln\left(1 + \frac{s_n^2}{\bar{x}^2}\right)} = \sqrt{\ln\left(1 + \frac{18.37}{3.63^2}\right)} = \sqrt{\ln(2.40)} = \sqrt{0.87} = 0.94$$

$$\hat{\sigma}_g(X) = e^{\hat{\sigma}} = e^{0.94} = 2.55$$

It can be observed that the estimates by the method of moments deviate slightly from the population parameters, affecting the calculation of the geometric mean, variance, and standard deviation. Therefore, these geometric statistics deviate from their population values.

$$X \sim Lognormal(\mu = 0.7, \sigma^2 = 1.28)$$

$$G(X) = \mu_g(X) = e^{\mu} = e^{0.7} = 2.01$$

$$GV(X) = \sigma_g^2(X) = E(\ln(X) - E[\ln(X)]) = e^{\sigma^2_{\ln(X)}} = e^{\sigma^2} = e^{1.28} = 3.60$$

$$GSD(X) = \sigma_g(X) = \sqrt{E(\ln(X) - E[\ln(X)])} = e^{\sigma_{\ln(X)}} = e^{\sigma} = e^{1.13} = 3.10$$

Maximum likelihood estimation and calculation of geometric mean, variance and standard deviation:

$$\hat{\mu} = \overline{\ln(x_i)} = \frac{\sum_{i=1}^{n} \ln(x_i)}{n} = \frac{21}{30} = 0.7$$

$$Mdn(X) = G(X) = \mu_g(X) = e^{\hat{\mu}} = e^{0.7} = 2.01$$

$$\hat{\sigma}^2 = s^2_{ln(x)} = \frac{\sum_{i=1}^{n}\left(\ln(x_i) - \overline{\ln(x)}\right)^2}{n} = \frac{38.4}{30} = 1.28$$

$$\hat{\sigma}^2 = e^{1.28} = 3.60$$

$$\hat{\sigma} = s_{\ln(x)} = \sqrt{\frac{\sum_{i=1}^{n}\left(\ln(x_i) - \overline{\ln(x)}\right)^2}{n}} = \sqrt{\frac{38.4}{30}} = \sqrt{1.28} = 1.13$$

$$\sigma_g(X) = e^{\hat{\sigma}} = e^{1.13} = 3.10$$

The maximum likelihood estimation was completely accurate without applying the bias correction. Confidence interval at 95% for the location parameter μ:

$$P\left(\overline{\ln(x)} - z_{1-\frac{\alpha}{2}}\sqrt{s^2_{\ln(x)}/n} \le \mu \le \overline{\ln(x)} + z_{1-\frac{\alpha}{2}}\sqrt{s^2_{\ln(x)}/n}\right) = 1 - \alpha$$

$$P\left(0.7 - 1.96 \times \sqrt{1.28/30} \le \mu \le 0.7 + 1.96 \times \sqrt{1.28/30}\right) = 0.95$$

$$P(\mu \in [0.30, 1.10]) = 0.95$$

Confidence interval at 95% for the geometric mean:

$$P\left(e^{\hat{\mu}_{\ln(x)} -\, {}_{1-\alpha/2}t_{n-1}\frac{\hat{\sigma}_{\ln(x)}}{\sqrt{n}}} \le \mu_g(X) = e^{\mu_{\ln(X)}} \le e^{\hat{\mu}_{\ln(x)} +\, {}_{1-\alpha/2}t_{n-1}\frac{\hat{\sigma}_{\ln(x)}}{\sqrt{n}}}\right) = 1 - \alpha$$

$$P\left(e^{0.7 - 2.05 \times \frac{1.13}{\sqrt{30}}} \le e^{\mu_{\ln(X)}} \le e^{0.7 + 2.05 \times \frac{1.13}{\sqrt{30}}}\right) = 0.95$$

$$P\left(e^{0.28} \le \mu_g(X) \le e^{1.12}\right) = P\left(\mu_g(X) \in [1.32, 3.07]\right) = 0.95$$

Calculation of the probability of having an average of at least eight sexual relations per week using parameters obtained through the maximum likelihood method:

$$P(X \ge x) = P\left(Z \ge \frac{\ln(x) - \hat{\mu}_{\ln(X)}}{\hat{\sigma}_{\ln(X)}}\right) = 1 - P\left(Z < \frac{\ln(x) - \hat{\mu}_{\ln(X)}}{\hat{\sigma}_{\ln(X)}}\right)$$

$$P(X \ge 8) = P\left(Z \ge \frac{\ln(8) - 0.7}{1.13}\right) = 1 - P(Z < 1.22) = 1 - 0.89 = 0.11$$

From the simulated example, 11 out of 100 women have an average of eight sexual relations per month (two per week).

Calculation of the density of having an average of eight sexual relations per month:

$$f_X(x = 8) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln(x) - \hat{\mu})^2}{2\sigma^2}} = \frac{1}{8 \times \sqrt{2 \times \pi \times 1.28}} e^{-\frac{(\ln(8) - 0.7)^2}{2 \times 1.28}} = 0.02$$

$$f_X(x = 8) = \frac{\varphi\left(\frac{\ln(8) - 0.7}{1.28}\right)}{8} = \frac{\frac{e^{-\frac{(\ln(8) - 0.7)^2}{2 \times 1.28}}}{\sqrt{2 \times \pi \times 1.28}}}{8} = \frac{0.17}{8} = 0.02$$

What average number of sexual relations has a cumulative probability of 90%?

$$x_{(p=0.9)} = e^{0.7 + 1.13 \times \Phi^{-1}(p=0.9)} = e^{0.7 + 1.13 \times 1.28} = e^{2.15} = 8.58$$

$$\Phi^{-1}(p = 0.9) = Z_{p=0.9} = 8.58$$

In Figure 4, the probability density function *f*X(*x*) and cumulative distribution function *F*X(*x*) are plotted with the 30 sample data. The calculations used in these plots are shown in the last two columns of Table 1.

It is important to assess the fit of the empirical data to a lognormal distribution. To achieve this, a dual strategy can be employed. Firstly, the inferential approach using Anderson and Darling's goodness-of-fit test is adopted (Table 2), as recommended by Neamvonk and Phuenaree (2022). Additionally, a graphical approach involves plotting the theoretical quantiles against the empirical quantiles, known as a quantile-quantile plot (Figure 5).

Statistical hypotheses for Anderson-Darling goodness-of-fit test"

H$_0$: X ~ *Lognormal*(μ, σ^2) and H$_1$: X ≁ *Lognormal* (μ, σ^2).

Testing statistic to assess the fit to a lognormal distribution with unknown parameters estimated from the sample:

$$z_{y_i} = \frac{y_i - m_y}{s_y} = \frac{\ln(x_i) - \overline{\ln(x)}}{s_{\ln(x)}} = \frac{\ln(x_i) - 0.7}{1.13}$$

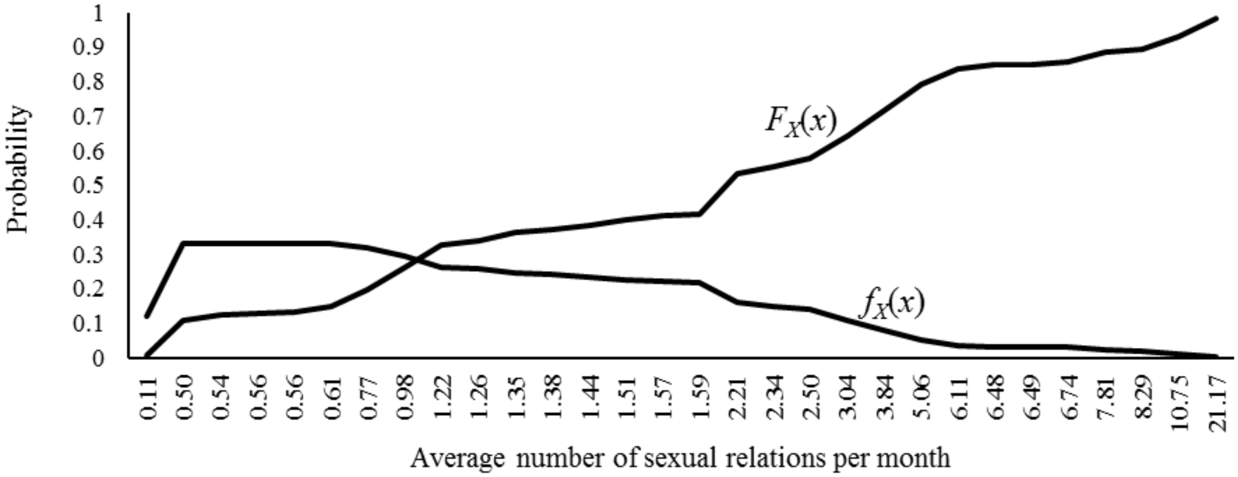

Figure 4. Plot of the probability density function $f_X(x)$ and cumulative distribution function $F_X(x)$ with the 30 sample data. Source: Author's elaboration.

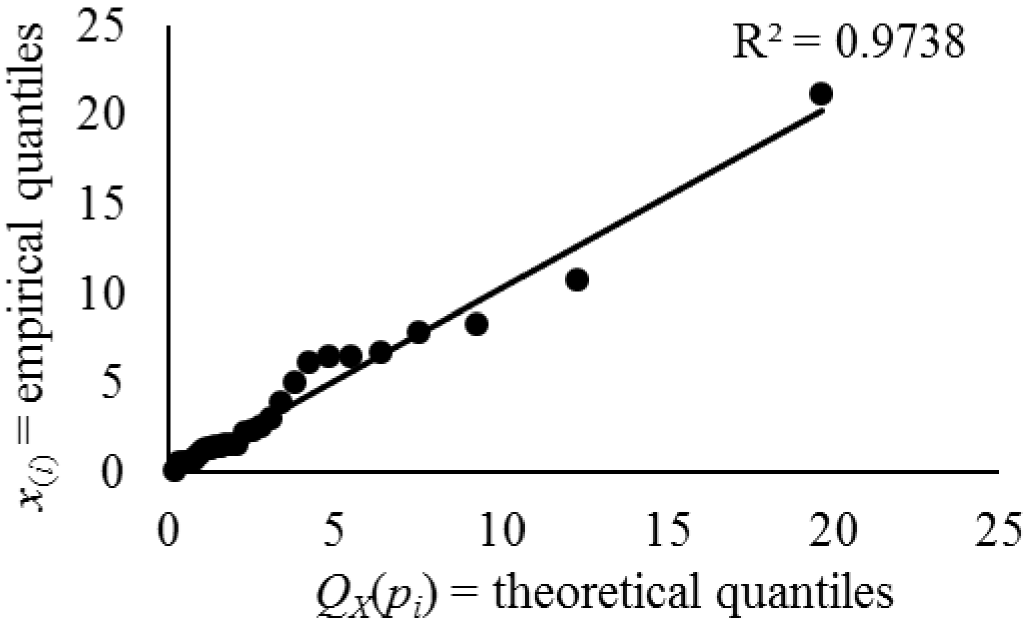

Figure 5. Quantile-quantile (q-q) plot with theoretical distribution: Lognormal (μ = 0.7, σ^2 = 1.28). Source: Author's elaboration.

$$AD = -n - \sum_{i=1}^{n}\left(\frac{2i-1}{n}\left(\ln\left[\Phi\left(z_{y_i}\right)\right] + \ln\left[1-\Phi\left(z_{y_{n+1-i}}\right)\right]\right)\right) = -30 + 30.36 = 0.36$$

$$AD_c = AD\left(1 + \frac{0.75}{n} + \frac{2.25}{n^2}\right) = 0.36\left(1 + \frac{0.75}{30} + \frac{2.25}{30^2}\right) = 0.37$$

The test statistic follows the Anderson-Darling distribution: ADc ~ A^2.

Decision based on the critical value or quantile of order 1 – α of the Anderson-Darling distribution A^2 for an α = 0.05 (Zaiontz, 2023).

$$AD_c \leq A_{1-\alpha}^2 = a_{1-\alpha}\left(1 - \frac{b_{1-\alpha}}{n} - \frac{d_{1-\alpha}}{n^2}\right) \Rightarrow H_0 \ se \ mantiene$$

$$AD_c > A_{1-\alpha}^2 \Rightarrow H_0 \ se \ rechaza$$

$$AD_c = 0.37 < A_{.95}^2 = 0.75\left(1 - \frac{0.795}{30} - \frac{0.89}{30^2}\right) = 0.73, se \ mantiene \ H_0$$

Table 2. Anderson-Darling test.

i	x_i	$y_i = \ln(x_i)$	$(2i-1)/30$	$z_i = z_{yi}$	$\Phi(z_i)$	z_{n+1-i}	$1-\Phi(z_{n+1-i})$	$\ln[\Phi(z_i)] + [1-\Phi(z_{n+1-i})]$	S_i
1	0.111	-2.199	0.033	-2.562	0.005	2.079	0.019	-9.234	-0.308
2	0.498	-0.697	0.1	-1.235	0.108	1.480	0.069	-4.890	-0.489
3	0.544	-0.608	0.167	-1.156	0.124	1.250	0.106	-4.337	-0.723
4	0.555	-0.588	0.233	-1.138	0.127	1.198	0.115	-4.219	-0.985
5	0.565	-0.572	0.3	-1.124	0.131	1.067	0.143	-3.982	-1.194
6	0.613	-0.489	0.367	-1.051	0.147	1.034	0.151	-3.813	-1.398
7	0.772	-0.259	0.433	-0.847	0.198	1.033	0.151	-3.509	-1.521
8	0.978	-0.022	0.5	-0.638	0.262	0.981	0.163	-3.154	-1.577
9	1.217	0.196	0.567	-0.445	0.328	0.814	0.208	-2.686	-1.522
10	1.259	0.231	0.633	-0.415	0.339	0.570	0.284	-2.339	-1.481
11	1.349	0.299	0.7	-0.354	0.362	0.364	0.358	-2.045	-1.431
12	1.382	0.324	0.767	-0.333	0.370	0.191	0.424	-1.852	-1.420
13	1.442	0.366	0.833	-0.295	0.384	0.132	0.447	-1.762	-1.468
14	1.513	0.414	0.9	-0.253	0.400	0.083	0.467	-1.677	-1.509
15	1.571	0.452	0.967	-0.219	0.413	-0.210	0.583	-1.423	-1.375
16	1.587	0.462	1.033	-0.210	0.417	-0.219	0.587	-1.408	-1.455
17	2.211	0.793	1.1	0.083	0.533	-0.253	0.600	-1.141	-1.255
18	2.339	0.850	1.167	0.132	0.553	-0.295	0.616	-1.078	-1.257
19	2.498	0.916	1.233	0.191	0.576	-0.333	0.630	-1.014	-1.251
20	3.041	1.112	1.3	0.364	0.642	-0.354	0.638	-0.892	-1.159
21	3.836	1.344	1.367	0.570	0.716	-0.415	0.661	-0.749	-1.024
22	5.060	1.621	1.433	0.814	0.792	-0.445	0.672	-0.630	-0.904
23	6.112	1.810	1.5	0.981	0.837	-0.638	0.738	-0.481	-0.722
24	6.479	1.869	1.567	1.033	0.849	-0.847	0.802	-0.385	-0.603
25	6.486	1.870	1.633	1.034	0.849	-1.051	0.853	-0.322	-0.526
26	6.736	1.907	1.7	1.067	0.857	-1.124	0.869	-0.294	-0.500
27	7.813	2.056	1.767	1.198	0.885	-1.138	0.873	-0.259	-0.458
28	8.286	2.115	1.833	1.250	0.894	-1.156	0.876	-0.244	-0.447
29	10.747	2.375	1.9	1.480	0.931	-1.235	0.892	-0.187	-0.355
30	21.170	3.053	1.967	2.079	0.981	-2.562	0.995	-0.024	-0.048
Σ									-30.362

i = order when data are ordered increasing, x_i = score in X (average number of sexual intercourse per month) ordered in ascending order, $y_i = \ln(x_i)$ = log-transformed (natural-based) x_i score, $(2i - 1) / 30$ = first factor in the product S_i, $z_i = z_{yi} = (y_i - m_y) / s_y$ = y_i score standardized using the mean and unbiased standard deviation of the Y sample data (m_y = 0. 7, $s_n(y)$ = 1. 1314), $\Phi(z_i)$ = cumulative probability up to z_{yi} in the standard normal distribution $N(0, 1)$, z_{n+1-i} = standardized y_i score ordered downward, $1 - \Phi(z_{n+1-i})$ = complement of the cumulative probability up to z_{n+1-i} in the standard normal distribution $N(0, 1)$, $\ln[\Phi(z_i)] + \ln[1 - \Phi(z_{n+1-i})]$ = sum of the natural logarithm of $\Phi(z_i)$ and the natural logarithm of $1 - \Phi(z_{n+1-i})$ or second factor in the product S_i, $S_i = (2i - 1) / 30) \times (\ln[\Phi(z_i)] + \ln[1- \Phi(z_{n+1-i})])$.
Source: Author's elaboration.

Decision based on the critical level or probability value with a significance level (α) of 5% (Zaiontz, 2023).

$$P(A^2 \le AD_c) = p = \begin{cases} 1 - e^{-13.436 + 101.14 \times AD_c - 223.73 \times AD^2} & AD_c \le 0.2 \\ 1 - e^{-8.318 + 42.796 \times AD_c - 59.938 \times AD_c^2} & 0.2 < AD_c \le 0.34 \\ e^{0.9177 - 4.279 \times AD_c - 1.38 \times AD_c^2} & 0.34 < AD_c < 0.6 \\ e^{1.2937 - 5.709 \times AD_c + 0.0186 \times AD_c^2} & AD_c \ge 0.6 \end{cases}$$

$$P(A^2 \le AD_c) \ge \alpha \Longrightarrow H_0 \textit{ is accepted.}$$

$$P(A^2 \le AD_c) < \alpha \Longrightarrow H_0 \textit{ is rejected.}$$

$$P(A^2 \le AD_c) = e^{0.9177 - 4.279 \times AD_c - 1.38 \times AD_c^2} = e^{0.9177 - 4.279 \times 0.372 - 1.38 \times 0.372^2} = 0.42$$

$$P(A^2 \le AD_c) = 0.42 < \alpha = 0.05, H_0 \textit{ is accepted}: X \sim \textit{Lognormal}(\mu, \sigma^2).$$

The theoretical quantiles $Q_X(p_i)$ for the quantile-quantile (q-q) plot are listed in the last column of Table 1 and are arranged on the abscissa axis. These theoretical quantiles are used to predict the empirical quantiles $x_{(i)}$, found in the second column of Table 1 and are placed on the ordinate axis of the q-q plot. The quantile order is computed using the median of the order statistics *i* from a sample of size *n* drawn from a standard uniform distribution *U* [0, 1], following the guidelines of Hyndman and Fan (1996). The calculation for the first data point $x_{(1)} = 0.11$ can be seen as follows:

$$p_i = \left((i) - 1/3\right)/(n + 1/3) = (1 - 1/3)/(30 + 1/3) = 0.02$$

$$Q_X(p_1 = 0.02) = e^{\mu+\sigma\times\Phi^{-1}(p_i)} = e^{0.7+1.13\times\Phi^{-1}(p_1=0.02)} = e^{0.7+1.13\times-2.01} = e^{-1.58}$$
$$= 0.21; X \sim Lognormal(\mu = 0.7, \sigma^2 = 1.28)$$

$$(Q_X(p_1 = 0.02) = 0.21, \quad x_{(1)} = 0.11)$$

The points align closely around the line of central tendency in the q-q plot, as supported by a multiple correlation or shared variance of 0.97 between the theoretical and empirical quantiles (Figure 5). This provides evidence of a strong fit of the sample data to the theoretical model.

Conclusions

From the illustrated presentation of the lognormal distribution with parameters μ and σ^2 (in logarithmic scale), it is evident that the calculation of probabilities and descriptive measures is straightforward. The probability calculations are simplified due to the relationship of this asymmetric and leptokurtic distribution with the normal distribution. Additionally, when computing descriptive measures, geometric measures are distinguished from arithmetic measures concerning the mean, variance, standard deviation, and coefficient of variation.

It is important to note that the arithmetic standard deviation is the square root of the arithmetic variance, but this relationship does not hold for the geometric standard deviation and variance. The most reliable estimators are those derived from the maximum likelihood method, providing direct parameters on a logarithmic scale.

The multiplicative effect of independent multiple factors on a variable follows a lognormal distribution, analogous to the linear effect of independent multiple factors on a variable that leads to a normal distribution. Owing to the connection between these two distributions, the multiplicative effect can be transformed into an additive one through lognormal transformation (Haines et al., 2020).

Naturally, it is essential to assess the fit of empirical data to the probability model using both graphical tools (histogram, frequency polygon, and quantile-quantile plot) and inferential methods (Anderson-Darling test). Finally, it is emphasized that this distribution can serve as a suitable probability model for certain variables related to sexual behavior, resource distribution, risk management, reaction times, physiological measurements, and epidemiology, such as predicting COVID-19 deaths. Therefore, it can be used to plan public health policy.

REFERENCES

Aggrawal N, Arora A, Anand A, Dwivedi Y (2021). Early viewers or followers: a mathematical model for YouTube viewers' categorization. Kybernetes 50(6):1811–1836.

Ahundjanov BB, Akhundjanov SB (2019). Gibrat's law for CO_2 emissions. Physica A: Statistical Mechanics and its Applications 526: article 120944.

Akhundjanov SB, Toda AA (2020). Is Gibrat's "economic inequality" lognormal? Empirical Economics 59:2071–2091.

Al-Masri AQS (2022). On combining independent tests in case of log-normal distribution. American Journal of Mathematical and Management Sciences 41(4):350–361.

Alasmar M, Clegg R, Zakhleniuk N, Parisis G (2021). Internet traffic volumes are not Gaussian – They are log-normal: An 18-year longitudinal study with implications for modelling and prediction. IEEE/ACM Transactions on Networking 29(3):1266-1279.

Andersson A (2021). Mechanisms for log normal concentration distributions in the environment. Scientific Reports 11(1):16418.

Aslam M (2024). Generating imprecise data from log-normal distribution. Measurement: Interdisciplinary Research and Perspectives, pp. 1-9.

Balci M, Kumral M (2024). Impacts of grade distribution and economies of scale on cut-off grade and capacity planning. Mining, Metallurgy and Exploration 1-23.

Balthrop AT (2021). Gibrat's law in the trucking industry. Empirical Economics 61(1):339–354.

Beykaei S, Abekah J, Rahim A (2020). Integration of uncertainty in profit planning: a current application. Journal of Applied Mathematics and Computation 4(4):195–205.

Chalkias A, Xenos M (2022). Relationship of effective circulating volume with sublingual red blood cell velocity and microvessel pressure difference: a clinical investigation and computational fluid dynamics modeling. Journal of Clinical Medicine 11(16):4885.

Chan P, Marchand M, Yoshida K, Vadhavkar S, Wang N, Lin A, Wu B, Ballinger M, Sternheim N, Jin JY, Bruno R (2021). Prediction of overall survival in patients across solid tumors following atezolizumab treatments: A tumor growth inhibition–overall survival modeling framework. CPT: Pharmacometrics and Systems Pharmacology 10(10):1171–1182.

Chen J, Korsunsky AM (2021). Why is local stress statistics normal, and strain lognormal? Materials and Design 198:109319.

Choi G, Buckley JP, Kuiper JR, Keil AP (2022). Log-transformation of independent variables: must we? Epidemiology 33(6):843–853.

Ciccone A (2021). Gibrat's law for cities: evidence from World War I casualties. CESifo Working Paper No. 9006:3830211.

Cobb CW, Douglas PH (1928). A theory of production. American Economic Review 18(1):139−165.

Eguia JX, Xefteris D (2022). Lognormal (Re) Distribution: A Macrofounded Theory of Inequality. University of Cyprus Working Papers in Economics.

Elassaiss-Schaap J, Duisters K (2020). Variability in the log domain and limitations to its approximation by the normal distribution. CPT: Pharmacometrics and Systems Pharmacology 9(5):245–257.

Finlay WH, Darquenne C (2020). Particle size distributions. Journal of Aerosol Medicine and Pulmonary Drug Delivery 33(4):178−180.

Fossion R, Sáenz-Burrola A, Zapata-Fonseca L (2020). On the stability and adaptability of human physiology: Gaussians meet heavy-tailed distributions. Inter Disciplina 8(20):55–81.

Galton F (1879). The geometric mean in vital and social statistics. Proceedings of the Royal Society of London 29(A):365–367.

Gayathri M, Malar MC, Manickam A, Pandey PM (2022). A mathematical model for prolactin secretion in healthy adults using lognormal distribution. International Journal of Health Sciences 6(S5):1536–1543.

Gechert S, Havránek T, Havránková Z, Kolcunova D (2019). Death to the Cobb-Douglas production function. Institut für Makroökonomie und Konjunkturforschung Working Paper.

Gibrat R (1931). Les inégalités économiques [Economic inequalities]. Paris: Librairie du Recueil Sirey.

Goldenholz DM, Westover MB (2023). Flexible realistic simulation of seizure occurrence recapitulating statistical properties of seizure diaries. Epilepsia 64(2):396–405.

González-Val R (2021). The Spanish spatial city size distribution. Environment and Planning B: Urban Analytics and City Science 48(6):1609–1631.

Gualandi S, Toscani G (2019). Human behavior and lognormal distribution. A kinetic description. Mathematical Models and Methods in Applied Sciences 29(4):717−753.

Guo Y, Li Z (2021). A lognormal model for evaluating maximum residue levels of pesticides in crops. Environmental Pollution 278:116832.

Haines N, Kvam PD, Irving LH, Smith C, Beauchaine TP, Pitt MA, Ahn WY, Turner B (2020). Theoretically informed generative models can advance the psychological and brain sciences: lessons from the reliability paradox. PsyArXiv:xr7y3.

Hyndman RJ, Fan Y (1996). Sample quantiles in statistical packages. The American Statistician 50(4):361–365.

Ito H, Shigeta K, Yamamoto T, Morita S (2022). Exploring sexual contact networks by analyzing a nationwide commercial-sex review website. PLoS ONE 17(11):e0276981.

Jokhadze V, Schmidt WM (2020). Measuring model risk in financial risk management and pricing. International Journal of Theoretical and Applied Finance 23(02):2050012.

Kapteyn JC (1903). Skew frequency curves in biology and statistics. Astronomical Laboratory and Popko Noordhoff, Groningen.

Kapteyn JC, van Uven MJ (1916). Skew frequency curves in biology and statistics. Hoitsema Brothers, Groningen.

Kawuki J, Sserwanja Q, Mukunya D, Sepenu AS, Musaba MW (2021). Prevalence and factors associated with sexual violence among women aged 15–49 years in rural Uganda: evidence from the Uganda Demographic and Health Survey 2016. Public Health 196(1):35–42.

Kirkwood TBL (1979). Geometric means and measures of dispersion. Biometrics 35(4):908–909.

Lahcene B (2021). Probability distributions related to modeling epidemic spread data “COVID-19 status and developments”. Journal of Applied Mathematics and Computation 5(2):134−144.

Lauritzen S, Uhler C, Zwiernik P (2019). Maximum likelihood estimation in Gaussian models under total positivity. Annals of Statistics 47(4):1835−1863.

Lawrence FTE, Cynthia OU, Oyinebifun BE (2023). Log normal distribution approach for forecasting the spread of Covid-19 in Nigeria. American Journal of Applied Mathematics and Statistics 11(1):13-21.

Maccone C (2020). Life expectancy and life energy according to Evo-SETI theory. In: Evo-SETI. Life Evolution Statistics on Earth and Exoplanets 2020:253-278.

Major CG, Paz-Bailey G, Hills SL, Rodriguez DM, Biggerstaff BJ, Johansson M (2021). Risk estimation of sexual transmission of Zika virus – United States, 2016–2017. The Journal of Infectious Diseases 224(10):1756–1764.

Materu J, Konje ET, Urassa M, Marston M, Boerma T, Todd J (2023). Comparison of survival analysis approaches to modelling age at first sex among youth in Kisesa Tanzania. Plos One 18(9):e0289942.

McAlister D (1879). The law of the geometric mean. Proceedings of the Royal Society of London 29(196-199):367-376.

McAloon C, Collins Á, Hunt K, Barber A, Byrne AW, Butler F, Casey M, Griffin J, Lane E, McEvoy D, Wall P, Green M, O'Grady L, More SJ (2020). Incubation period of COVID-19: a rapid systematic review and meta-analysis of observational research. BMJ Open 10(8):e039652.

Minyu F, Liang-Jian D, Feng C, Matjaž P, Jürgen K. (2020). The accumulative law and its probability model: an extension of the Pareto distribution and the log-normal distribution. Proceedings of the Royal Society of London 476A(2237):20200019.

Neamvonk J, Phuenaree B (2022). Assessment of Anderson-Darling and their modified tests for right skewed distribution. Computer Science 17(3):1327−1339.

Olosunde AA, Ejiofor C (2021). Log-exponential power distribution for accelerated failure time model in survival analysis and its application. Afrika Statistika 16(1):2587-2603.

Özkan R, Sen F, Balli S (2020). Evaluation of wind loads and the potential of Turkey's south west region by using log-normal and gamma distributions. Wind and Structures 31(4):299–309.

Pal R, Huang Z, Yin X, Lototsky S, De S, Tarkoma S, Liu M, Crowcroft J, Sastry N (2021). Aggregate cyber-risk management in the IoT age: Cautionary statistics for (re)insurers and likes. IEEE Internet of Things Journal 8(9):7360–7371.

Phelps CE (2023). Optimal health insurance. Journal of Risk and Insurance 90(1):213–241.

Ranger J, Kuhn JT, Ortner TM (2020). Modeling responses and response times in tests with the hierarchical model and the three-parameter lognormal distribution. Educational and Psychological Measurement 80(6):1059–1089.

Rao DS, Rajasekhar C, Naidu G (2022). A study on the statistical distribution and regression analysis of novel coronavirus in India. Journal of Medical Pharmaceutical and Allied Sciences 11(3):4888−4894.

Sinharay S, van Rijn PW (2020). Assessing fit of the lognormal model for response times. Journal of Educational and Behavioral Statistics 45(5):534–568.

Smirnov RG, Wang K. (2021). The Cobb-Douglas production function revisited. In DM Kilgour, H Kunze, R Makarov, R Melnik, X Wang (Eds.), Recent developments in mathematical, statistical and computational sciences. AMMCS 2019. Springer Proceedings in Mathematics and Statistics 343:725–734. Cham: Springer International Publishing.

Stuart A, Ord K (2010). Kendall's advanced theory of statistics. Vol. 1. Distribution theory (6th ed.). John Wiley & Sons, New York.

Swat MJ, Grenon P, Wimalaratne S (2016). ProbOnto: ontology and knowledge base of probability distributions. Bioinformatics 32(17):2719−2721.

Valvo PS (2020). A bimodal lognormal distribution model for the prediction of COVID-19 deaths. Applied Sciences 10(23):8500.

Verbavatz V, Barthelemy M (2020). The growth equation of cities. Nature 587(7834):397–401

Wang J, Gottschal P, Ding L, van Veldhuizen DA, Lu W, Houssami N, Greuter MJW, de Bock GH (2021). Mammographic sensitivity as a function of tumor size: A novel estimation based on population-based screening data. The Breast 55(1):69–74.

Waymyers S, Chakraborty H (2024). Modeling Negatively Skewed Survival Data in Accelerated Failure Time and Correlated Frailty Models. Journal of the Indian Society for Probability and Statistics 25(3):343–371.

Wikipedia Contributors (2023). Log-normal distribution. In Wikipedia (Ed.), The free encyclopedia. https://en.wikipedia.org/w/index.php?title=Log-normal_distribution&oldid=1137817833

Zaiontz C (2023). One-sample Anderson-Darling test. In Real Statistics using Excel. https://real-statistics.com/non-parametric-tests/goodness-of-fit-tests/anderson-darling-test/

Zijian L (2020). A theorem on a product of lognormal variables and hybrid models for children’s exposure to soil contaminants. Environmental Pollution 263(B):A114393.