International Journal of Medical Research and Innovation
2025, Vol. 1, No. 1, pp. 14–19
https://doi.org/10.11648/j.ijmri.20250101.12

Methodology Article

Geometric Statistical Measures in the Analysis of Skewed and Zero-Valued Data: Implications for Biomedical Research

Jesus Gomez[1, *], Tibisay Rangel[1], Victor Gomez[2]

[1]School of Pharmacy, Central University of Venezuela, Caracas, Venezuela
[2]School of Mathematics, Central University of Venezuela, Caracas, Venezuela

Abstract

Research in fields such as biomedicine often generates data that do not conform to a normal distribution, exhibiting positive skewness (right skewness), as is characteristic of the log-normal distribution. In these scenarios, the use of the arithmetic mean (AM) and standard deviation (SD) can lead to misinterpretations of central tendency and dispersion, as the AM is sensitive to extreme values and overestimates wide numerical ranges. This paper presents a guide on the use of the Geometric Mean (GM), the Geometric Standard Deviation (GSD), and the Geometric Coefficient of Variation (GCV) as the most appropriate statistical tools for this type of data, as well as for sets with disparate numerical scales. The fundamental practical challenge of zero values, common in biomedical measurements such as viral loads or analyte concentrations, is addressed. A specific methodology for treating data with zeros is detailed and justified, consisting of the addition and subsequent subtraction of a unit to allow the calculation of geometric statistics. Finally, two approaches for calculating the Geometric Coefficient of Variation are analyzed and compared, highlighting its nature as a power basis rather than a simple mathematical ratio, and discussing its comparative utility despite its complex interpretation. The need for a better understanding and application of these metrics to improve the accuracy and reproducibility of data analyses is emphasized.

Keywords

Geometric Mean, Geometric Standard Deviation, Geometric Coefficient of Variation, Log-normal Distribution

1. Introduction

The application of the geometric mean, the geometric standard deviation and the geometric coefficient of variation in biomedical research has gained increasing importance in data analysis, especially in contexts where data normality cannot be assumed.

The geometric mean (GM) is a measure of central tendency primarily used for data sets with a logarithmic distribution. Unlike the arithmetic mean, which can be influenced by outliers, the GM is less sensitive to these extremes, making it a useful tool in biomedical research, where data are often multiplicative or exponential in nature [1].

On the other hand, the geometric standard deviation (GSD) is used to measure the variability of data on a logarithmic scale. This measure is particularly useful in studies where the

*Corresponding author: jalbgomez@gmail.com (Jesus Gomez)

Received: 7 September 2025; **Accepted:** 30 October 2025; **Published:** 8 December 2025

data are distributed asymmetrically, allowing for a better interpretation of the dispersion compared to the arithmetic standard deviation [2]. The geometric coefficient of variation (GCV) is defined as the ratio of the GSD to the GM, providing a relative measure of variability that is independent of the scale of the data [3].

Although the arithmetic mean and standard deviation are the most commonly used metrics in biomedical research, their applicability is limited in many contexts. Data that present skewed distributions or include outliers can lead to misinterpretations when using these traditional metrics. It is crucial to consider the properties of the population distribution and how they are represented in the sample. To describe the results of the study, the most appropriate summary statistics must be chosen, taking into account the specific question and the assumptions of the data analysis. The "average" is often reported without adequate consideration of its implications. Ignoring these aspects can lead to inconsistencies in the calculation of variance and confusion about summary statistics [4]. The geometric mean and its derivatives offer a robust alternative that can improve the accuracy of statistical analyses in these cases.

Despite its advantages, the geometric mean is not suitable for all data types. For example, it cannot be calculated for negative or zero values, which limits its applicability in certain contexts. Furthermore, the interpretation of the geometric mean may be less intuitive for some researchers, which may lead to its underutilization compared to the arithmetic mean [5].

The choice of the appropriate metric should be based on the nature of the data and the specific research questions. The geometric mean, geometric standard deviation, and geometric coefficient of variation are valuable statistical tools in biomedical research. Their application can improve data interpretation in contexts where traditional metrics are insufficient.

In data analysis, the arithmetic mean (AM) and standard deviation (SD) can lead to biased conclusions, especially when the data do not meet the assumption of normality. A vast number of biological variables, such as survival times, antibody titers, drug concentrations, parasite burden, or viral load, exhibit a positively skewed distribution. In these distributions, most values are concentrated at the lower end of the scale, with a "tail" extending toward higher values, often approximating a log-normal distribution.

Geometric measures (Geometric Mean, Geometric Standard Deviation and Geometric Coefficient of Variation) offer a much more robust conceptual and practical framework for addressing these challenges. Authors such as Kirkwood [3-6] have worked on formalizing these measures, but their application remains limited. This work aims to clarify the calculation methodology and correct interpretation of these metrics, contextualizing them within the geometric statistical framework and proposing a solution for handling zero values.

2. Fundamentals of Geometric Measurements

Geometric measures are designed for data whose changes are multiplicative or proportional in nature, not additive. Their power lies in their direct connection to the logarithmic transformation, which converts a log-normal distribution to a normal distribution. The geometric framework operates at higher hierarchical levels (powering and root extraction) compared to the arithmetic framework (addition and division), making it inherently more sensitive to dispersion and low values.

2.1. Geometric Mean (GM)

The GM is the measure of central tendency for data following a log-normal distribution. Conceptually, it represents the central value on a multiplicative scale. It is calculated as the nth square root of the product of all the values:

$$GM = (x_1 * x_2 * \ldots * x_n)^{1/n} = (\Pi_{i=1}^{n} x_i)^{1/n}$$

A fundamental property is that the logarithm of the geometric mean is equal to the arithmetic mean of the logarithms of the data:

$$ln(GM) = (1/n) * \Sigma_{i=1}^{n} ln(x_i)$$

This demonstrates that the GM is the antilogarithm of the arithmetic mean of the log-transformed data. For this reason, it is not unduly influenced by extreme values and provides a better estimate of central tendency in data with positive skewness. The GM tends to be smaller than the AM, and this difference becomes more pronounced as the dispersion increases.

2.2. Geometric Standard Deviation (GSD)

Unlike the arithmetic standard deviation (additive measure), the GSD is a multiplicative factor that describes how the data is scattered around the geometric mean. It is calculated as the antilogarithm of the standard deviation of logarithmically transformed data: $GSD = exp((ln(s))$ Where $(ln(s))$ is the standard deviation of the natural logarithms of the data:

$$(ln(s)) = \sqrt{[(\Sigma_{i=1}^{n} (ln(x_i) - ln(GM))^2) / (n)]}$$

The dispersion range is expressed as a multiplicative interval:

$$[GM / GSD; GM * GSD]$$

This interval will contain approximately 68% of the data in a log-normal distribution. For a log-normal distribution, the dispersion range containing approximately 68% of the data is expressed multiplicatively using the Geometric Mean (GM) and the Geometric Standard Deviation (GSD). Unlike the normal distribution, where the data are distributed symmetrically around the arithmetic mean, the log-normal distribution is asymmetric. In the latter, the logarithm of the data follows a normal distribution. While in a normal distribution the mean interval ± 1 standard deviation contains about 68% of the data, in a log-normal distribution a multiplicative interval is used. This parallelism is due to the fact that, when applying the logarithm to the variable, the distribution is transformed into a normal one, where addition and subtraction are natural operations. When you return to the original scale, these operations become multiplication and division.

2.3. The Inherent Link to the Log-Normal Distribution

The elegance of these measurements lies in their perfect alignment with the theory of log-normal distribution. If a random variable X follows a log-normal distribution, its logarithmic transformation, Y = ln(X), follows a normal distribution. The GM of X corresponds to the antilogarithm of the arithmetic mean of Y, and the GSD of X corresponds to the antilogarithm of the standard deviation of Y. By using GM and GSD, the principles of statistics are applied to normal distributions over the transformed data, then projecting the results back to the original scale, which is the conceptually correct approach for data of this nature.

3. The Problem and Handling of Zero Values

A critical obstacle to the direct application of geometric statistical measurements is the presence of values equal to zero. Since the logarithm of zero is indefinite, and GM implies the product of all values (which would be zero if one of them is), standard formulas cannot be calculated. In biomedicine, zeros are informative and frequent (e.g., undetectable viral load, absence of bacteria); Ignore or delete them it would seriously bias the results. Therefore, a standardized methodology is required for its management.

3.1. Proposed Methodology for the Management of Zeros

A simple and reversible transformation method is proposed to allow the calculation of geometric statistics in the presence of zeros. The methodology consists of adding a constant to all the values in the dataset before the calculation, and subtracting it from the final results. The most commonly used constant is 1. [7].

3.2. Justification

Zero Handling: By adding 1, the lowest possible value (0) becomes 1 (x' = 0 + 1 = 1). The logarithm of 1 is 0, which allows the data to be included in the calculations without generating a product of zero or an undefined logarithm.

Preservation of Relative Order: The addition of a constant does not alter the order or relative dispersion of the data on the logarithmic scale.

Reversibility: Subtracting 1 from the final measurements (GM and GSD) adjusts the result to reflect a magnitude closer to the original scale of the data.

3.3. Calculation Procedure

Transform the Data: For each xi observation in the dataset, calculate $x'_i = x_i + 1$.

1) Calculate the Modified Geometric Mean (GM $_{(mod)}$):

First, calculate the geometric mean of the transformed data (MG'):

$$GM' = \left(\Pi_i^{=1\ldots n} (x_i + 1)\right)/n$$

Then, subtract 1 to get the final measurement:

$$GM\_mod = GM' - 1$$

Calculate the Modified Geometric Standard Deviation (GSD_mod):

First, calculate the geometric standard deviation of the transformed data (GSD'):

$$GSD' = exp(\sqrt{}[(\Sigma_{i=1}^{n} (ln(x_i + 1) - ln(GM'))^2) / (n)])$$

Then, subtract 1 to get the final measurement:

$$GSD\_mod = GSD' - 1$$

This methodology allows for robust and defensible inclusion of zero values, providing stable estimates of central tendency and dispersion. It is crucial that this procedure is transparently reported in the methods section of any publication.

4. Geometric Coefficient of Variation (GCV)

The Coefficient of Variation (CV) is a standardized measure of the dispersion of a distribution, useful for comparing variability between different data sets. For log-normal data, several formulations have been proposed for its geometric analogue.

4.1. Lee Humphries' Proposal

Humphries proposes a direct analogy with the arithmetic formula (CV = SD / $\bar{X}$). The GCV is calculated by reducing the geometric standard deviation to the power of the reciprocal of the geometric mean: [8].

$$GCV\ (Humphries)\ =\ GSD^{\ (1\,/\,GM)}$$

The interpretation of this measure is less direct than that of its arithmetic counterpart, since it does not represent a simple reason. Its meaning is revealed in the inverse relationship: the GCV is the base that, raised to the power of the geometric mean, results in the geometric standard deviation $(GCV)^{(GM)} = DesvGeom$. This formulation, although mathematically coherent, is not very intuitive and makes it difficult to use practically in isolation, but its true value emerges in the comparative analysis of multiple data series.

4.2. Proposal Based on Logarithmic Variance (SAS/Wicklin)

This approach, popularized in the statistical literature and used in software such as SAS, defines the GCV directly from the variance of logarithmically transformed data:

$$GCV\ (Wicklin)\ =\ \sqrt{}[exp(((ln(s))^2)\ -\ 1]$$

where *((ln(s))* ²is the variance of the natural logarithms of the original data [9].

Advantages:

Direct Connection: This formula is directly derived from the theoretical properties of the log-normal distribution.

Interpretation: For small values, the result of this formula approaches the standard deviation of the log-transformed data ($s_{(ln(x)}$), which facilitates its interpretation as a measure of relative variability on the logarithmic scale.

For most biomedical applications, Wicklin's formula is the most recommended because of its strong theoretical foundation and wide acceptance.

4.3. Confidence Intervals

It is important to distinguish between two methods for calculating intervals:

1) Method 1: Logarithmic transformation + classical interval.

The confidence interval is calculated in the logarithmic (normalized) space and then transformed back to the original space with the exponential function $exp()$. This method is statistically more rigorous if the log-transformed data follow a normal distribution, allowing the sample size to be incorporated and the interval to be adjusted precisely.

2) Method 2: Multiplication/Division by Geometric Standard Deviation $[GM\ /\ GSD;\ GM\ *\ GSD]$.

This method generates a range around the geometric mean by multiplying/dividing by the GSD. It is very intuitive and easy to calculate, useful for representing relative dispersion on multiplicative scales. However, it does not represent a confidence interval in the classical statistical sense and does not incorporate the sample size or confidence level. (Based on the proposal of Humphries, L. (2010) for Geometric Coefficient of Variation [8]).

5. The Logarithmic Transformation: A Key Point to Avoid Errors in the Analysis of Biased Data

It is very important to note that when data is not normally distributed, there is a risk of obtaining misleading or erroneous results. Therefore, data transformations, such as logarithmic transformations, are used to adjust the frequency distribution of the data to a more normal shape. This adjustment is vital, as it allows for the reliable application of parametric tests, improving the interpretation and quality of the results obtained [10].

The logarithmic transformation is a widely used statistical technique [11-13]. In the context of linear regression, data transformation is commonly used in the health field when data are skewed or do not fit a normal distribution, which is the distribution assumed for linear regression residuals [14].

The logarithmic transformation is a convenient means of transforming a highly skewed variable into one that is more approximately normal. This technique helps to convert nonlinear relationships, which often arise in many applications [15], in linear relationships, which can reduce data bias [11]. An example of this occurs with data that have a strong right bias, where the logarithmic transformation can help to achieve a more symmetric distribution, closer to the normal one [12].

Statistical tests and models are developed using logarithmically transformed data. It is important to note that tests performed with logarithmically transformed data often help in understanding the behavior of the original, untransformed data [11]. To achieve a practical interpretation, it is necessary to apply the inverse function of the logarithm, which is the exponential function.

It is essential that data transformations be applied with great caution [14]. Although the logarithmic transformation is one of the most popular transformations when dealing with skewed data, its popularity has also made it vulnerable to misuse, which can lead to incorrect interpretations and errors [12].

The logarithmic transformation is a statistical tool, but its arbitrary use may be inappropriate [12]. Caution is necessary because once the variable is transformed, the interpretation becomes more complex, and the transformation could create artificial patterns or reduce the accuracy of the model if the underlying relationship is not exponential or multiplicative.

Furthermore, if there is a lack of adjustment of the transformed variable, the logarithmic transformation can cause more problems than it solves, as it tends to distort the relationships between the variables.

The logarithmic transformation is an essential tool in biomedical research, but its application must be guided by a thorough understanding of the data and the objectives of the analysis. If the transformed variable is not properly fitted, this technique can create more problems than it solves, distorting the relationships between variables. Recognizing and correctly handling lognormal distributions not only facilitates the interpretation of results but also improves their communication, making them clearer and more effective. Proper handling of these distributions is crucial to avoid misinterpretations and ensure the validity of the findings [16].

6. Conclusions and Recommendations for Practice

The correct characterization of biomedical data is a fundamental pillar for the validity of any research. Reflection on the geometric statistical framework reveals powerful but underutilized tools.

Distribution Diagnosis: Before applying any measures, researchers must evaluate the distribution of their data. If a log-normal (or positively asymmetry) distribution is suspected, the use of GM and GSD is imperative.

Zero Management: The methodology of adding and subtracting unity is a practical and defensible approach to including zero values in geometric calculations. It is crucial that this procedure is transparently reported in the methods section of any publication.

Choice of GCV: The use of the formula is recommended $\mathrm{GCV} = \sqrt{[\exp(\ln(s)^2) - 1]}$ (Wicklin) to evaluate relative variability, due to its theoretical robustness and its widespread implementation in statistical packages. Understanding that the GCV conceptually measures the deviation relative to the average (using a 'power base' for its calculation) is to its correct application.

Results Reporting: When publishing, it should be clearly specified that the Geometric Mean and Geometric Standard Deviation are being reported. The dispersion interval must be presented in its multiplicative form (e.g., GM = 5.0, GCV = 1.5; with a range of [3.33, 7.5]) in addition to the classical interval.

Pedagogical Gap: A significant gap in standard pedagogical material on this topic is identified and the need to develop more educational resources that facilitate the understanding and application of these metrics is suggested.

Future Research: Future research could focus on the application and validation of GCV in fields such as medicine and biology, exploring its usefulness in areas such as epidemiology, genetics, and pharmacology. This will consolidate its relevance as a standard statistical tool in biomedical research. The adoption of these practices will increase the accuracy of statistical analyses and improve the reproducibility and correct interpretation of research results.

Abbreviations

AM	Arithmetic Mean
CV	Coefficient of Variation
GCV	Geometric Coefficient of Variation
GM	Geometric Mean
GSD	Geometric Standard Deviation
GCV	Geometric Coefficient of Variation
$MG_{(mod)}$	Modified Geometric Mean
SD	Standard Deviation

Conflicts of Interest

The authors declare that they have no competing interests. All contributing authors certify that they have no financial or non-financial affiliations (such as personal or professional relationships, academic competition, or intellectual beliefs) that could be perceived as influencing the objectivity of the research presented. The authors have no financial or proprietary interests in any material, product, or organization mentioned in this manuscript.

References

[1] Vogel RM. The geometric mean? Commun Stat Theory Methods. 2020; 50(16): 3833-41. https://doi.org/10.1080/03610926.2020.1743313

[2] Troon John Benedict, Karanjah Anthony, Alilah Anekeya David. Modeling Geometric Measure of Variation About the Population Mean. American Journal of Theoretical and Applied Statistics. Vol. 8, No. 5, 2019, pp. 179-184. https://doi.org/10.11648/j.ajtas.20190805.13

[3] Kirkwood, T. B. L. (1979). Geometric means and measures of dispersion. *Biometrics*, 35(4), 908-909.

[4] Martinez MN, Bartholomew MJ. What Does It "Mean" ? A Review of Interpreting and Calculating Different Types of Means and Standard Deviations. Pharmaceutics. 2017 Apr 13; 9(2): 14. https://doi.org/10.3390/pharmaceutics9020014

[5] De La Cruz, R., & Kreft, J. U. (2018). Geometric mean extension for data sets with zeros. arXiv preprint arXiv: 1806.06403.

[6] Kirkwood, T. B. L. (1993). Geometric standard deviation reply to Bohidar. *Drug Development and Industrial Pharmacy*, 19(3), 395-396.

[7] Brandling-Bennett AD, Anderson J, Fuglsang H, Collins R. Onchocerciasis in Guatemala. Epidemiology in fincas with various intensities of infection. *Am J Trop Med Hyg.* 1981 Sep; 30(5): 970-81.

[8] Humphries, L. (2010). The Geometric Coefficient of Variation. *ThinkingApplied.com. Mind Tools: Applications and Solutions*.

[9] Wicklin, R. (April 27, 2011). Compute the geometric mean, geometric standard deviation, and geometric CV in SAS. The DO Loop. https://blogs.sas.com/content/iml/2019/10/02/geometric-mean-deviation-cv-sas.html

[10] Habibzadeh F. Data Distribution: Normal or Abnormal? J Korean Med Sci. 2024; 39: e35. https://doi.org/10.3346/jkms.2024.39.e35

[11] Hospitality. Institute. Log Transformation in Regression: Achieving Linearity. 2024 Mar 6. Disponible en: (Fuente sin URL proporcionada en los extractos).

[12] Yonchev M. Log Transformations and their Implications for Linear Regression. Medium. 2023 Mar 16. Disponible en: (Fuente sin URL proporcionada en los extractos).

[13] Feng C, Wang H, Lu N, Chen T, He H, Lu Y, et al. Log-transformation and its implications for data analysis. Shanghai Arch Psychiatry. 2014; 26(2): 105-8. https://doi.org/10.3969/j.issn.1002-0829.2014.02.009

[14] Jones L, Barnett A, Vagenas D. Linear regression reporting practices for health researchers, a cross-sectional meta-research study. PLoS ONE. 2025; 20(3): e0305150. https://doi.org/10.1371/journal.pone.0305150

[15] Benoit K. Linear Regression Models with Logarithmic Transformations. London: Methodology Institute, London School of Economics; 2011 Mar 17.

[16] Motulsky HJ, Head T, Clarke PBS. Analyzing lognormal data: A nonmathematical practical guide. Pharmacol Rev. 2025; 77: 100049. https://doi.org/10.1016/j.pharmr.2025.100049