Asian Journal of Probability and Statistics

Volume 24, Issue 4, Page 23-34, 2023; Article no.AJPAS.106212
ISSN: 2582-0230

Determinants of Estimate Difference between Geometric Measure and Standard Deviation

Rukia Mbaita Mbaji [a], Troon John Benedict [b] and Okumu Otieno Kevin [a*]

[a] Department of Mathematics and Physical Sciences, Maasai Mara University, Narok, Kenya.
[b] Department of Economics, Maasai Mara University, Narok, Kenya.

Authors' contributions

This work was carried out in collaboration among all authors. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/AJPAS/2023/v24i4532

Open Peer Review History:
This journal follows the Advanced Open Peer Review policy. Identity of the Reviewers, Editor(s) and additional Reviewers, peer review comments, different versions of the manuscript, comments of the editors, etc are available here: https://www.sdiarticle5.com/review-history/106212

Original Research Article

Received: 04/06/2023
Accepted: 08/09/2023
Published: 16/09/2023

Abstract

Measures of variation are statistical measures which assist in describing the distribution of data set. These measures are either used separately or together to give a wide variety of ways of measuring variability of data. Researchers and mathematicians found out that these measures were not perfect, they violated the algebraic laws and they possessed some weakness that they could not ignore. As a result of these facts, a new measure of variation known as geometric measure of variation was formulated. The new measure of variation was able to overcome all the weaknesses of the already existing measures. It obeyed all the algebraic laws, allowed further algebraic manipulation and was not affected by outliers or skewed data sets. Researchers were also able to determine that geometric measure was more efficient than standard deviation and that its estimates were always smaller than those of standard deviation but they did not determine their main relationship and how the sample characteristics affect the minimum difference between geometric measure

**Corresponding author: Email: kevinotieno15@gmail.com;*

and standard deviation. The main aim of this study was to empirically determine the ratio factor between standard deviation and geometric measure and specifically how certain variable such as sample size, outliers and geometric measure affects the minimum difference between geometric measure and standard deviation. Data simulation was the concept that was used to achieve the studies objectives. The samples were simulated individually under four different types of distributions which were normal, Poisson, Chi-square and Bernoulli distribution. A Hierarchical linear regression model was fitted on the normal, skewed, binary and countable data sets and results were obtained. Based on the results obtained, there is always a positive significant ratio factor between the geometric measure and standard deviation in all types of data sets. The ratio factor was influenced by the existence of outliers and sample size. The existence of outliers increased the difference between the geometric measure and standard deviation in skewed and countable data sets while in binary it decreased the difference between the standard deviation and geometric measure. For normal and binary data sets, increase in sample size did not have any significant effect on the difference between geometric measure and standard deviation but for skewed and countable data sets the increase in sample size decreased the difference between geometric measure and standard deviation.

Keywords: Geometric; standard deviation; Data simulation; distribution; hierarchical regression.

1 Introduction

Variation of data is the key characteristics of data sets that tells us how data sets are distributed. It is usually very important to measure the variation of every data set since measures of variation describes the distribution of data sets, defines the width of the data set and how the data values are spread out from each other and the central tendency. Measure of variations consists of four type and they include; range, standard deviation, variance, geometric measurer and interquartile range. These measures can be used separately or together so that they can give a wide variety of ways of measuring the variability of data and each measure of variation have different functions. Therefore, this study investigated the determinants of estimate difference between geometric and standard deviation.

1.1 Standard Deviation

As a measure of variation, standard deviation measures how much the data values deviate from the mean or the closeness of a particular data from the mean. Standard deviation also tends to be small and show little variation when the data is closely concentrated to the mean while its larger when the data are spread out from the mean hence showing more variation. Data is described the best when mean is paired with the standard deviation. This type of measure of variation does not violate any algebraic laws, therefore it allows further algebraic manipulations that gives estimates that are of the same unit as the initial data set but it is affected by skewed data sets and outliers [1,7,13]. Standard deviation is computed as:

$$s = \sqrt{\frac{\sum (X_i - \bar{X})}{n-1}} \tag{1}$$

$$\sigma = \sqrt{\frac{\sum (X_i - \mu)}{N}} \tag{2}$$

Previous research confirmed that measure of variations contained some weaknesses. The research determined that range as a measure of variation is influenced by outliers and does not give any information about the distribution values, variance gives estimates that are not the same as the initial datasets therefore skewing the data more while standard deviation is affected by skewed data sets or data sets with outliers. These weaknesses could not be ignored by the researchers or even the statisticians. [9,12,10] research and was able to model a new measure of variation called the geometric measure of variation, this measure of variation was able to overcome all the weakness of the already existing measures of variation. Geometric measure of variation was able to allow further algebraic manipulation and not violate any algebraic laws because it focused on the product about the mean rather than the sum. It was also not affected by the skewed data sets and outliers.

The research of [9,10] also determined that geometric measure of variation gave estimates that were smaller than those of the standard deviation. However, this study did not determine how these estimates were compared to the standard deviation, the main relationship between geometric measure and standard deviation, whether sample size and existence of the outlier affected the relation or even how these measures were related to each other. Given the shortcomings of the past research, this study aimed at empirically investigating the main relation between standard deviation and geometric measure and how the existence of outlier and sample size of the data set affected this relationship. The study was also interested in determining the ratio factor that relates between geometric measure to standard deviation.

1.2 Geometric measure of variation

Standard deviation as a measure of variation has mostly been used because it is capable of allowing further algebraic manipulations to be carried out. The algebraic relationship between standard deviation and other measures of variation is; Standard deviation is mainly used together with the mean in describing the kind of distributions, it basically measures the distance of the data values from the mean. While with the interquartile range, it is usually preferred over range because the outliers tend to impact the range more than how they impact standard deviation and SD can be calculated from all data [2,17]. Despite the fact that standard deviation and variance both help in finding the distribution of data in a population, SD gives more clarity about the deviation from the mean and it is related to variance because it is the square root of variance [22].

Past studies have determined that the measure of variations including the standard deviation have the following weaknesses; range does not give any information about the distribution of the observations from the measure of central tendency, variance gives result that are not the same as the initial datasets therefore skewing the data more and standard deviation is always affected by skewed data sets and data sets with outliers [16,18]. Therefore, there was a gap and need for researchers to develop a new measure of variation that will solve these weaknesses.

The study by [10], developed a model of a measure of variation that was able to overcome the weaknesses of the current measures by not violating the algebraic laws, giving the estimates which are not of the same unit as the initial datasets. Geometric measure of variation about the mean was formulated to estimate the average variations about the mean for discrete and continuous, weighted and unweighted data sets. Based on this research, some conclusions were made. It was determined that the geometric measure gave estimates that were smaller than the estimates of standard deviate on and these estimates and averages of the measure of variation were not affected by outliers or skewed datasets and allows further algebraic manipulations to be carried on. The comparison of their efficiency was also determined using bias, mean square error or relative efficiency methods and if the results are less than 1 then the SD is more efficient, if it is more than 1 then geometric measure is more efficient and if it is equal to 1 then the efficiency of geometric measure and standard deviation is equal. Due to the fact that geometric, peaked and skewed data sets are highly affected by the outliers, past studies successfully determined that geometric measure is more efficient than the standard deviation in estimating the average variations. Its main focus is on the product of the absolute deviation about the mean and can be used on weighted and unweighted data sets, discrete and continuous data sets.

A good measure of variation from the mean should be able to obey all the algebraic laws, previous studies formulated a geometric measure of variation about the mean which did not violate the three algebraic laws, so in the pursuit of determining the relationship between geometric measure and standard deviation some basic concepts were introduced: Arithmetic mean is simply the mean or the average, it is the simplest and the most widely spread used measure of a mean. Arithmetic mean is determined by summing all the observational values and dividing them by the number of the observation in the data set [19,21]. It is determined as,

$$\overline{X} = \frac{\sum_{i=1}^{n} Xi}{n} \tag{3}$$

We also have geometric averaging which usually gives averages that are not affected by outliers and also ones that do not assume symmetry of datasets and it uses a theory to get the geometric mean. Therefore, geometric mean is the mean that indicates the central tendency or typical values of a set of values by using their products, it differs from the arithmetic mean because it takes into account the compounding that occurs from period to period [20]. It is computed as:

$$GM = n\sqrt{\prod_{i=1}^{n} C_i} \tag{4}$$

Past studies found that, compared to other means such as harmonic and arithmetic mean, the geometric mean gives results that are greater than the harmonic mean and smaller than the arithmetic mean when the data values are positive or greater than 0. Therefore, it can only be used to determine the average of positive numbers. We have standard deviation as an interesting measure of variation which is the average squared distance from the mean and it is calculated as the square root of variance by determining each data point's deviation relative to the mean [7,21]. Previously we stated that geometric measure focused on the product of absolute deviation about the mean. The computation of both the standard deviation and geometric measure depended on the different types of datasets such as the weighted and unweighted datasets and discrete and continuous data sets.

Consider a data with vectors $V = [V_1, V_2, V_3, ..., V_n]$, to be a set of unweighted data sets. Hence the standard deviation and geometric measure of this data set from the mean $\overline{d}$ is computed respectively as

$$\left|\overline{d}\right| = \sqrt[n]{\prod_{i=1}^{p} \left|d_i\right|} \tag{5}$$

$$G_{\mu} = \begin{cases} \exp\left(\dfrac{1}{n}\sum_{i=1}^{p} \ln\left(\left|d_i\right|\right)\right) & \forall d_i \neq 0 \\ o \quad \text{elsewhere} & \forall d_i = 0 \end{cases} . \tag{6}$$

While for weighted datasets, let's consider vectors of unweighted data sets $V = [V_1, V_2, V_3, ..., V_n]$ and vectors of weights $\varsigma = \left[\varsigma_1, \varsigma_2, \varsigma_3, ...\varsigma_n\right]$ .Then a new vector of weighted data $\varsigma^V = \left[\varsigma_1 v_1, \varsigma_2 v_{2,} \varsigma_3 v_{3,} .....\varsigma_n v_n\right]$ has a standard deviation and geometric measure respectively as

$$\left|\overline{d}\right| = \sqrt[\sum_{i=1}^{n} \varsigma_i]{\prod_{i=1}^{p} \left|d_i\right|^{\varsigma_i}} \tag{7}$$

$$G_w = \begin{cases} \exp\left(\dfrac{1}{\sum_{i=1}^{n} \varsigma_i} \sum_{i=1}^{p} \varsigma_i \ln\left(\left|d_i\right|\right) \right) & \forall d_i \neq 0 \\ 0 \quad \text{elsewhere} & \forall d_i = 0 \end{cases} \quad . \tag{8}$$

For discrete data sets, these are data sets that can only take on certain values [10,17].

So, let’s assume that the variable v_1 is discrete with probability mass function $\varsigma(v_i) \quad \forall = 1,2......n$ and O otherwise then geometric is formulated as

$$G_{pm} = \begin{cases} \exp\left(\sum_{i=1}^{n} \varsigma(v_i) \bullet \ln\left|d_i\right| \right) & \forall d_i \neq 0 \\ & \forall d_i = 0 \end{cases} \tag{9}$$

While continuous datasets are quantitative data that represent a scale of measurement that can consist of numbers other than whole numbers [9,10,19]. They consist of continuous variable, assume that v is continuous on an interval $a \leq v \leq b$ with probability density function $\varsigma(v)$.

$$G_{pd} = \begin{cases} \exp\left(\int_{a}^{b} \varsigma(v) \bullet \ln\left|d\right| dd \right) & \forall d_i \neq 0 \\ 0 \quad \text{elsewhere} & \forall d_i = 0 \end{cases} \tag{10}$$

The study by [9,10], formulated a model of geometric measure in terms of efficiency, the geometric measure was found to be more efficient than standard deviation in estimating average measures of variation, through various test he found out that geometric measure is always smaller than the standard deviation.

2 Methods

2.1 Data Simulation

Data simulation was the main backbone in the study’s methodology. We used simulation, where one sets the ground rules of a random process then the computer uses the random numbers to generate an outcome that adhered to the rules while data simulation took a large amount of data and used it in mirroring conditions in the real world to the determine the best method of validating a model [10]. The R program was used in simulating data. This program came up with a set of pseudo random numbers generators from different distributions that allowed the simulation of data from these distributions. Since our methodology was mainly interested in simulation, the R program used the “r” generator function.

Data of different sample sizes with n observations from different types of distributions were simulated. The total number of samples used was 100 samples and each sample had different sample sizes. Each of the samples had four different variables namely; standard deviation, geometric measure, outliers and sample sizes. Any type of data always had two types of variables, independent and dependent variables. Independent variable is a type of variable whose variation does not depend on another variable and is always controlled to test its effects on other variables while dependent variables are variables whose values and variations depend on another variable [5,11]. In the study standard deviation was the dependent variable being tested while geometric measure, outliers and sample sizes were the independent variables whose effects were tested to see how they affected and controlled the standard deviation. Observations with different sample sizes were simulated and they either had outliers or no outliers. Normal, Bernoulli, Poisson and Chi-square distribution were the different distributions that were used in the data simulation. The reason as to why the study was only interested on using these four types of distributions is because they represented the most common and possible data sets in life since we always expect data sets to be either normal, dummy, countable or even skewed [5,19]. Let us see how they individually impacted our study and how data from the respective distributions were simulated;

2.2 Normal Distribution

This is the most common distribution whereby it represented data that was normal or not normal. This distribution was always assumed to have a mean of 0 and a standard deviation of 1. In simulation it used the “r” generator function rnorm (), and the total simulation function was rnorm (n, mean=0, SD=1) where n was the sample size or the number of observations in a sample [6].

2.3 Bernoulli distribution

This was the type of distribution that represented data with two possible outcomes. The variables with only two probabilities of success or failure are called the dummy or binary variables. These variables took values of 0 or 1, indicating that there was a possible outcome of success or failure. The trials without 0 or 1 outcomes were said to have outliers. The function rbern (n, prob) was the function used in simulating Bernoulli distribution where prob was the probability [3,6].

2.4 Chi-square distribution

This distribution was a representation of skewed data sets. This was proven by the curve of the data sets. The curve of the chi-square distribution was always seen to be skewed to the right and the degree of freedom directly affected the shape of the curve, an increase in the degrees of freedom increased the shape of the curve making it more symmetrical. The degree of freedom denoted as df was always equal to the distribution mean and it was twice the variance [9,13]. The ‘r’ generator function used in simulation was rchisq (n, df)

2.5 Poisson distribution

This was a distribution of countable variables. It events are always independent and they occur in a fixed period of time. It was known as a special type of binomial distribution when n goes to infinity and the expected number of successes remains fixed. This distribution tends to describe the distribution of binary data from a very infinite sample. Used the function rpois (n, λ) in simulating thedata (“Poisson distribution,” n.d)

In our program, we had a table with four columns each representing the different variables; sample size, outliers, standard deviation and geometric measure. Simulation of the sample was done by coding different types of data with different sample sizes like 60,87,10 etc. But the sample sizes were simulated either with or without outliers. If they had outliers, the program recorded a one on the column of the outliers but if the data did not have any outliers it recorded a zero. Standard and geometric measure of the sample sizes were the calculated using the following formulas;

$$s = \sqrt{\frac{\sum (X_i - \overline{X})^2}{n-1}} \tag{11}$$

$$G_{pm} = \begin{cases} \exp\left(\sum_{i=1}^{n} \varsigma(\nu_i) \bullet \ln\left|d_i\right| \right) & \forall d_i \neq 0 \\ & \forall d_i = 0 \end{cases} \tag{12}$$

The results were then recorded in their respective columns, this was done continuously in different sample sizes from the four distribution and the samples added up to a total of 100 samples [15].

2.6 Determining the relationship

The relationship between the geometric measure and standard deviation was determined by the regression method. A linear regression model was formulated to effectively determine the relationship and see how the independent variables affected the standard deviation. Linear regression is a type of model that used a straight line to describe the relationship between variables. It consists of two types: simple and multiple linear regression. Simple linear regression used only one independent variable while multiple linear regression used more than independent variables [4,8]. The multiple regression models have certain assumptions that needed to be satisfied before doing anything first and they included; the observations should be independent, there should be no autocorrelation between the independent variables and this was checked using the function cor () and data should be normally distributed, this was checked using the function hist (). Since the data was simulated, the study made sure that the data used already satisfied the assumptions, so there was no need of checking again.

The study had three independent variables, so the multiple linear regression model was the best model used in determining the relationship. A multiple linear regression model with standard deviation as the dependent variables and outliers, sample sizes and geometric measure as the independent variables was fitted and the significant of each variable was tested. The model was of the form

$$y = a + b_1x_1 + b_2x_2 + b_3x_3 + \varepsilon_i \tag{13}$$

Where; y = standard deviation

a = minimum difference

x_1 =geometric measure

x_2 =sample size

x_3 = outliers

The significant of the model was tested to see how the independent variables affected the relationship and the following hypothesis was used:

H_0 : Variable is not significant

H_1 : Variable is significant

If the p-values of the independent variables are less the alpha value of 0.05 then we concluded that there was a significant relationship between the dependent and the independent variable. If b_1, b_2, b_3 had a p-value of less than α , we rejected the H_0 which was the null hypothesis and conclude that sample size, outliers and geometric measure significantly affected the relationship by the values of b respectively [4,15]. If the values of

b_1, b_2, b_3 were positive then we say that as the variables increases then the standard deviation also increases by the values of b's but the value of b_3 was the ratio factor between the geometric measure and standard deviation. This ratio factor enabled us to get the standard deviation through the geometric measure without even using it original formula ("R –multiple regression," n.d.)

2.7 Hierarchical linear regression

This is a type of linear regression that was used to compare successive regression models and determine the significance that each of the above variables had. This form of multiple linear regression was the best and we used it to determine the relationship and significance of the variables step by step [14,15]. We analysed our variables using a step-by-step method. This was accomplished by having several regression models, starting with the smallest order to the largest, we started with a simple model and went by increasing and adding the independent variables step by step on the previous model. Our interest was to determine how the newly added variables behaved and how they show any significant improvements in the adjusted R^2. R^2 was the proportion of explained variance in the dependent variable by the model [16]. We looked at the adjusted R^2 for the successive models and found their differences. We had three successive models since our study had three independent variables. In each model we looked at the p-value of the independent variables, if the p-values are less than the alpha value then we reject the null hypothesis and conclude that the variables were significant and they affected the relationship. We closely looked at the value of the adjusted R^2, if the value increased when other variables are added to the previous model then we say that there was an improvement in R^2. The model with the highest value of adjusted R^2, was said to be the most efficient model. The most efficient model included the variables that were significant and affected the relationship. The variable that decreased the value of another variable when it was added to the previous model, it was said to be most effective variable, meaning that it was the variable that affected and explained the standard deviation or the relationship more than the other variables. This analysis enabled us to determine how and by what value, the independent variables affected the standard deviation and its relationship with the geometric measure.

3 Results and Discussion

Several samples with different sample sizes were simulated under different distributions and they include normal, Bernoulli, Chi-square and Poisson distributions.

3.1 Normal distribution

Under this distribution 100 samples with different sample sizes were simulated using the function rnorm (n, mean, SD) and they either had outliers or no outliers, whereby 50 samples had outliers and the rest had no outliers. The 100 sample characteristics (outliers, sample size, standard deviation and geometric measure) were used to fit the respective hierarchical model. The results are illustrated in the Table 1.

Table 1. Hierarchical regression on normal distribution

	Dependent variables			
Standard deviation				
	(1)	**(2)**	**(3)**	**(4)**
Geometric	0.966***	0.968***	0. 676***	0.682***
N		-0.0001	-0.0002	
Outlier			2.997***	2.915***
Constant	2.550***	2.606***	3.364***	3.275***
Observations	100	100	100	100
R2	0.993	0.993	0.994	0.994
Adjusted R2	0.993	0.993	0.994	0.994
Residual std. Error	0.416 (df=98)	0.426 (df=9)	0.393(df=96)	0.395 (df=97)
Note			*p<0.1; **p<0.05; ***p<0.01	

Under normal distribution the results consisted of four models. In model 1 we only had two variables whereby geometric measure was the independent variable while standard deviation was the dependent variable. In this model geometric measure was significant with a ratio factor of 0.9660. When sample size was added in model 2 the geometric measure remained significant while the variable sample size negatively affected the relationship between geometric measure and standard deviation because its coefficient was insignificantly different from 0. The addition of outliers in model 3 affected the ratio factor and the relationship between geometric measure and standard deviation. The existence of outliers in the data sets increased the difference between geometric measure and standard deviation by 2.997 units on average. The variable sample size was insignificant from the model; it was therefore eliminated and only the outlier and geometric measure were included. This increased the ratio factor from 0.676 to 0.683. We conclude that only the existence of the outliers in the data sets influenced the difference between geometric measure and standard deviation, however, the sample size had no influence on the difference between the geometric measure and standard deviation for normal data sets.

3.2 Bernoulli distribution

10 samples were simulated with different sample size, where out of the 100, 50 were simulated with outliers and the rest without outliers. The study then used the simulated data to compare the respective standard deviation and geometric measure for each sample. The 100 sample characteristics (outliers, sample size, standard deviation and geometric measure) were used to fit a hierarchical model. Table 2 illustrates the results obtained.

Table 2. Hierarchical regression on Bernoulli distribution

	Dependent variables			
Standard deviation				
	(1)	**(2)**	**(3)**	**(4)**
Geometric	0.406***	0.406***	0. 398***	0.398***
N		0.0000	0.000	
Outlier			-0.001***	-0.001***
Constant	0.303***	0.303***	0.307***	0.307***
Observations	100	100	100	100
R2	0.994	0.994	0.995	0.995
Adjusted R2	0.994	0.994	0.995	0.995
Residual std. Error	0.001 (df=98)	0.001(df=97)	0.001(df=96)	0.001 (df=97)
Note			**p<0.1; **p<0.05; ***p<0.01*	

The Table 2 consists of four models, the first model with only the geometric measure as the independent variable, this model shows that the geometric measure had a significant ratio factor and the minimum difference between the geometric measure and standard deviation was 0.303. The variable sample size was then added in the next model, this variable did not have any significant effect on the ratio factor or the minimum difference between geometric measure and standard deviation they neither increased or decreased. The addition of outlier in the data set was determined to significantly affect the ratio factor and the relationship between geometric measure and standard deviation. They decrease the difference between geometric measure and standard deviation by 0.001 units on average. The variable sample size was eliminated from the fourth model due to the fact that sample size remained an insignificant contributor to the relationship between geometric measure and standard deviation. This did not affect the effects of the outliers, the ratio factor or the minimum difference because the coefficients neither increased nor decreased.

In conclusion, only the existence of the outliers in the data sets had an influence on the difference between geometric measure and standard deviation where it reduced the difference. However, the sample size had no influence on the difference between the geometric measure and standard deviation for binary data sets.

3.3 Chi-square distribution

Samples with or without outliers were simulated with different sample sizes using the function rchisq (n, df). Standard deviation and geometric measure were simulated using different functions and formulas. A hierarchical model was fitted using the sample characteristics (outliers, sample size, standard deviation and geometric measure) from the 100 samples and the results are shown in Table 3.

Table 3. Hierarchical regression on Chi-square distribution

	Dependent Variable		
Standard Deviation			
	(1)	**(2)**	**(3)**
Geometric	1.312***	1.312***	0. 895***
N		-0.0002**	-0.0002**
Outlier			0.385***
Constant	0.861***	0.967***	1.364***
Observations	100	100	100
R2	0.532	0.551	0.674
Adjusted R2	0.527	0.541	0.664
Residual std. Error	0.316(df=98)	0.311(df=97)	0.266(df=96)
Note		**p<0.1; **p<0.05; ***p<0.01*	

The results in Table 3 shows the existence of three model. In the first model, geometric measure as the only independent variable was significant and it positively affected the relationship between geometric measure and standard deviation by increasing the minimum difference between the two variables by 0.861 units on average and the ratio factor was 1.312. The variable sample size was later added in model 2, the addition of this variable increased the minimum difference between geometric measure and standard deviation from 0.861 to 0.967 thus making the variable significant. For model 3, the existence of the variable outlier reduced the ratio factor from 1.312 to 0.895 but increased the minimum difference t0 1.364. The addition of the variable outlier was established to affect the ratio factor between the standard deviation and geometric measure.

In conclusion, the existence of both outliers and sample size in the data sets had an influence on the difference between geometric measure and standard deviation for skewed data sets.

3.4 Poisson distribution

The Poisson distribution used the function rpois (n, λ) to simulate 100 samples with different sample sizes. The study then computed the respective standard deviation and geometric measure and using the sample characteristics (outliers, sample size, standard deviation and geometric measure) a hierarchical model was fitted. The results are illustrated in Table 4.

Table 4. Hierarchical regression on poisson distribution

	Dependent Variable		
Standard Deviation			
	(1)	**(2)**	**(3)**
Geometric	1.065***	1.072***	0.577***
N		-0.0001	-0.0001*
Outlier			0.395***
Constant	0.787***	0.826***	1.066***
Observations	100	100	100
R2	0.565	0.572	0.820
Adjusted R2	0.561	0.563	0.814
Residual std. Error	0.216(df=98)	0.215(df=97)	0.141(df=96)
Note		**p<0.1; **p<0.05; ***p<0.01*	

Geometric measure was the only independent variable in model 1. This model shows that geometric measure had a positive significant ratio factor 0f 1.065and the minimum difference between the geometric measure and standard deviation was found to be 0.787. In model 2, the variable sample size was added, this increased the ratio factor and minimum difference between geometric measure and standard deviation from 1.0651 to 1.072 and from 0.787 to 0.826 respectively. Sample size did not significantly affect the ratio factor and the minimum difference between the geometric measure and standard deviation, because it coefficient was not significantly different from 0. For model 3, the variable outlier was then added, it affected the ratio factor between the standard deviation and geometric measure. The existence of the outlier in the data set was also established to increase the difference between the geometric measure and standard deviation by 0.395 units on average. However, the addition of the outliers, made sample size a significant contributor to the difference between geometric measure and standard deviation and decreasing the difference by 0.0001 units on average.

In conclusion, for countable data sets the existence of outlier and the sample size both had a significant contribution towards the difference between geometric measure and standard deviation in that the geometric measure is in most cases less than the standard deviation.

4 Conclusion

Based on the results obtained from the simulation done under different distributions. It was established that there is always a positive significant ratio factor between geometric measure and standard deviation. The effects that the sample characteristics had on the relationship between geometric measure and standard deviation varied under different distributions. For binary data sets, the difference between the geometric measure and standard deviation decreased due to the existence of outliers, while for normal, skewed and countable data sets. The existence of the outlier was found to increase the difference between the geometric measure and standard deviation. Increase in sample size was determined to decrease the difference between the geometric measure and standard deviation for skewed and countable data sets, however it had no significant effect on the difference between the geometric measure and standard deviation in normal and binary data sets.

Competing Interests

Authors have declared that no competing interests exist.

References

[1] Hargrave M. How to use standard deviation to measure risk. Investopedia; 2021.
Available:https://www.investopedia.com/terms/s/standarddeviation.asp

[2] Mondal S. Measures of variability*:* 5 Types statistics. Biology Discussion; 2016. Available:https://www.biologydiscussion.com/genetics/measures-of-variability-5-types-statistics/38125

[3] Abegyan M. Why is the standard deviation The most widely used measure of dispersion explain?; 2020.
Available:https://findanyanswer.com/why-is-the-standard-deviation-the-most-widely-used-measure-of-dispersion-explain

[4] Predamkar P, Linear Regression in R. How to intrepret Linear Regression with Examples. EDUCBA; 2019.
Available:https://www.educba.com/linear-regression-in-r/

[5] Troon JB, Anthony K, David A. Estimating average variation about the population mean using geometric measure of variation. International Journal of Statistical Distributions and Applications. 2020;6(2):23.

[6] Troon JB, Anthony K, David A. Modelling geometric measure of variation about the population mean. American Journal of Theoretical and Applied Statistics. 2019;8(5):179.

[7] Lawson JD, Lim Y. The geometric mean, matrices, metrics and more. The American Mathematical Monthly; 2001.

[8] Schuetter J. Chapter 1. In J. Schuetter, measurers of dispersion; 2007;45-54.

[9] Vidya. Basic statistics for exploring data: Measures of Variation. Daydreaming Numbers; 2019.
Available:https://daydreamingnumbers.com/blog/measures-of-variation/

[10] Raymondo J. Measures of variation from statistical analysis in the behavioral sciences. Kendall hunt publishing; 2015.

[11] Mindlin D. On the relationship between arithmetic and geometric returns. Cdi Advisors Research LLC; 2011.

[12] Roenfeldt K. Better than average: Calculating geometric means using SAS. Henry M. Foundation for the advancement of military medicine; 2018.

[13] Bhardwaj A. Comparative study of various measurers of dispersion. Journal of Advances in mathematics. 2013;1(1).

[14] Drew Robb. What is data simulation?. Datamation; 2021.

[15] Lee DK, In J, Lee S. Standard deviation and standard error of the mean. Korean Journal of Anesthesiology. 2015;68(3):220.

[16] Andrew D. StackPath; 2021.
Available:https://www.bodyloveconference.com/blog/what-is-the-relation-between-mean-and-standard-deviation/

[17] Todd Helmenstine. Understand the difference between independent and dependent variables. ThoughtCo; 2022.

[18] El Omda S, Sergent SR. Standard deviation. PubMed; StatPearls Publishing; 2021.
Available:https://www.ncbi.nlm.nih.gov/books/NBK574574/

[19] Bernoulli Distribution in R. Geeks for Geeks; 2021.

[20] Bevans R, Linear Regression in R. An easy step-by-step guide. Scribbr; 2020.
Available:https://www.scribbr.com/statistics/linear-regression-in-r/

[21] How To Calculate Standard Deviation (Plus Definition and Use). (n.d.). Indeed Career Guide.
Available:https://www.indeed.com/career-advice/career-development/how-to-calculate-standard-deviation

[22] Quick JM. R tutorial series: Hierarchical linear regression r-bloggers. R-Bloggers; 2010.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
https://www.sdiarticle5.com/review-history/106212