Original Research Article

DETERMINANTS OF ESTIMATE DIFFERENCE BETWEEN GEOMETRIC MEASURE AND STANDARD DEVIATION

Abstract: Measures of variation describes and analyses the distribution of data sets while variation of data usually tells how data is distributed. Measures of variation are important because they help researchers in defining the width of a data set and how the data values are spread out from each other. There are four types of measures of variation and they include; range, interquartile range, variance and standard deviation. 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 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 the geometric measure of variation was formulated and it was able to overcome all the weaknesses of the already existing measures. Geometric measure of variation obeyed all the algebraic laws, allowed further algebraic manipulation and was not affected by outliers or skewed data sets. This research determine that geometric measure of variation was more efficient than the standard deviation and it estimates were smaller than those of standard deviation but they did not determine the main relationship between the two measures of variations and how the sample characteristic; sample size, outliers and geometric measure affect the minimum difference between geometric measure and standard deviation.

Keywords: Geometric,Standarddeviation,Outliers,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. In general, these measures of variations are used to analyse the distribution of data points in the data sets. Measure of variations consists of four type and they include; range, standard deviation, variance and interquartile range. These measures of variations 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.

1.1 Variance

Display of the data is described by using variance, the larger the variance the more the data is spread out. As a measure of variation, variance is the average squared distance from the mean and measures how far each number in the data set is from the mean or from each other. Variance permits the partition of variations into the various components but alsodiffers from other measures because it compares every piece of value to the mean. Statisticians uses variance to compare every single data to another to see how they relate. Since it is a squared entity, variance gives no intuitive way to compare itself directly to the data values and it gives estimates that are not the same as the initial values in the data sets therefore making it not easily interpretable because it skews the data more [7,12]. Hence it is computed as:

$$s^2 = \frac{\sum\left(X - \bar{X}\right)}{n-1} (1)$$

$$\sigma^2 = \frac{\sum\left(X - \mu\right)}{N} (2)$$

1.2 Standard Deviation

As a measure of variation, standard deviationmeasures how much the data values deviates 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 [1,13]. Standard deviation is computed as:

$$s = \sqrt{\frac{\sum(X_i - \bar{X})}{n-1}} \quad (3)$$

$$\sigma = \sqrt{\frac{\sum(X_i - \mu)}{N}} \quad (4)$$

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. [9,10] 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 variationwas 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.

2. METHODS

Simulation of data was the main backbone of the study's methodology. Simulation is a process 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 takes a large around of data and used it in mirroring conditions in the real word to determine the best method of validating a model[5,11]. The R program and the 'r' generator function was used and it came up with a set pseudo random numbers generators from different distributions that allowed the simulation of data from these distributions.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 has two types of variables, independent and dependent variable: the independent variable is the variable whose variation does not depend on another variable while the dependent variable is the variable whose variations depends on another variable [2,15]. In the study the dependent variable was the standard deviation while the independent variables were geometric measure, sample size and outliers.

Observations with different sample sizes were simulated and they either had outliers or no outliers. These observations were simulated under Normal, Chi-square, Poison and Bernoulli distributions. These four distributions were used because they represented the most common and possible data sets in life. Normal, Chi-square, Poison and Bernoulli distribution represents data sets that are normal, skewed, countable and dummy respectively.

2.1 *Normal Distribution*

This distribution represents data that is normal thus making it the most common distribution. The distribution assumed a mean of 0 and a standard deviation of 1 and it used the 'r' generator function rnorm() to simulate the data [6].

2.2 *Bernoulli Distribution*

This was the distribution that represented dummy variables or data with two possible outcomes of success or failure and it used the function rbern (n, prob) to simulate the data. The 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[3,6].

2.3 *Chi-Square Distribution*

This distribution was a representation of skewed data sets and it used the generator function rchisq (n, df) for simulation whereby the degree freedom was denoted as df and it was equal to the distribution mean but twice the variance [5,7].

2.4 *Poisson Distribution*

Countable data sets were represented by the poison distribution. Countable data sets are data sets whose events are always independent and only occurs 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 remained fixed. The distribution used the function rpois (n, λ) in simulating the data.

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 do not have any outliers then it recorded a zero.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. Standard and geometric measure of the sample sizes were then calculated using the following formulas;

$$s = \sqrt{\frac{\sum (X_i - \overline{X})^2}{n-1}} \quad (5)$$

$$G_{pm} = \begin{cases} \exp\left(\sum_{i=1}^{n} \varsigma(\nu_i) \bullet \ln|d_i|\right) & \forall d_i \neq 0 \; (8)(6) \\ & \forall d_i = 0 \end{cases}$$

Regression method was used to determine the main relationship between the geometric measure and standard deviation. A linear regression model was fitted to see how the independent variables are affecting the standard deviation. There are two types of linear regression model, simple linear and multiple linear regression model. Simple regression model uses only one independent variable while multiple regression model uses more than one independent variables [4,8,15]. The study fitted a multiple linear regression model since it had more than one independent variable. 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 data should be normally distributed. Since the data was simulated, there was no need of checking if the assumptions were satisfied.

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

$$y = a + b_1 x_1 + b_2 x_2 + b_3 x_3 \; (7)$$

Where; $y =$ standard deviation

a = minimum difference

x_1 =geometric measure

x_2 =sample size

x_3 = outliers

The significance 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

Since a multiple linear regression was used in the determination of the relationship, a hierarchical model was specifically fitted. Hierarchical regression model is a 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, 16]. The variables were analysed using a step-by-step method. This was accomplished by having several regression models, starting with the smallest order to the largest. 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. The study 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.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& 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 it 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

The results in table3 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

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.

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.Theexistence 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.

References

[1] Abegyan, M. (2020, January 31). *Why is the standard deviation The most widely used measure of dispersion explain?* Findanyanswer.com. https://findanyanswer.com/why-is-the-standard-deviation-the-most-widely-used-measure-of-dispersion-explain

[2] Andrew, D. (2021, October 22). *StackPath*. Www.bodyloveconference.com. https://www.bodyloveconference.com/blog/what-is-the-relation-between-mean-and-standard-deviation/

[3] *Bernoulli Distribution in R*. (2021, April 15). GeeksforGeeks.

[4] Bevans, R. (2020, February 25). *Linear Regression in R | An Easy Step-by-Step Guide*. Scribbr. https://www.scribbr.com/statistics/linear-regression-in-r/

[5] Drew Robb. (2021, December 22). *What is data simulation?* Datamation

[6] El Omda, S., & Sergent, S. R. (2021). *Standard Deviation*. PubMed; StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK574574/

[7] Hargrave, M. (2021, April 15). *How to Use Standard Deviation to Measure Risk*. Investopedia. https://www.investopedia.com/terms/s/standarddeviation.asp

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

[9] Troon .J.B., Anthony, K., & David .A,. (2020). Estimating Average Variation About the Population Mean Using Geometric Measure of Variation. *International Journal of Statistical Distributions and Applications*, *6*(2), 23.

[10] Troon .J.B., Anthony, K., & David .A,.. (2019). Modelling Geometric Measure of Variation About the Population Mean. *American Journal of Theoretical and Applied Statistics*, *8*(5), 179

[11] Lee, D. K., In, J., & Lee, S. (2015). Standard deviation and standard error of the mean. *Korean Journal of Anesthesiology*, *68*(3), 220.

[12] Mondal, S. (2016, July 12). *Measures of Variability: 5 Types | Statistics*. Biology Discussion. https://www.biologydiscussion.com/genetics/measures-of-variability-5-types-statistics/38125

[13] Predamkar, P. (2019, September 2). *Linear Regression in R | How to intrepret Linear Regression with Examples*. EDUCBA. https://www.educba.com/linear-regression-in-r/

[14] Quick, J. M. (2010, January 15). *R Tutorial Series: Hierarchical Linear Regression | R-bloggers*. R-Bloggers.

[15] Todd Helmenstine. (2022, March 1). *Understand the Difference Between Independent and Dependent Variables*. ThoughtCo.

[16] Vidya. (2019, February 13). *Basic statistics for exploring data: Measures of Variation*. Daydreaming Numbers. https://daydreamingnumbers.com/blog/measures-of-variation/