American Journal of Theoretical and Applied Statistics
2019; 8(5): 179-184
http://www.sciencepublishinggroup.com/j/ajtas
doi: 10.11648/j.ajtas.20190805.13
ISSN: 2326-8999 (Print); ISSN: 2326-9006 (Online)

Modelling Geometric Measure of Variation About the Population Mean

Troon John Benedict[1], Karanjah Anthony[2], Alilah Anekeya David[3]

[1]Department of Mathematics and Physical Science, Maasai Mara University, Narok, Kenya

[2]Department of Mathematic, Multimedia University, Nairobi, Kenya

[3]Department of Mathematics and Statistics, Masinde Muliro University of Science and Technology, Kakamega, Kenya

Email address:

troonbenedict@gmail.com (T. J. Benedict), karanjah@mmarau.ac.ke (K. Anthony), aliladavid2010@gmail.com (A. A. David)

*Corresponding author

To cite this article:

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. doi: 10.11648/j.ajtas.20190805.13

Received: September 17, 2019; **Accepted**: September 28, 2019; **Published**: October 12, 2019

Abstract: Measures of dispersion are important statistical tool used to illustrate the distribution of datasets. These measures have allowed researchers to define the distribution of various datasets especially the measures of dispersion from the mean. Researchers and mathematicians have been able to develop measures of dispersion from the mean such as mean deviation, variance and standard deviation. However, these measures have been determined not to be perfect, for example, variance give average of squared deviation which differ in unit of measurement as the initial dataset, mean deviation gives bigger average deviation than the actual average deviation because it violates the algebraic laws governing absolute numbers, while standard deviation is affected by outliers and skewed datasets. As a result, there was a need to develop a more efficient measure of variation from the mean that would overcome these weaknesses. The aim of this paper was to model a geometric measure of variation about the population mean which could overcome the weaknesses of the existing measures of variation about the population mean. The study was able to formulate the geometric measure of variation about the population mean that obeyed the algebraic laws behind absolute numbers, which was capable of further algebraic manipulations as it could be used further to estimate the average variation about the mean for weighted datasets, probability mass functions and probability density functions. Lastly, the measure was not affected by outliers and skewed datasets. This shows that the formulated measure was capable of solving the weaknesses of the existing measures of variation about the mean.

Keywords: Standard Deviation, Geometric Measure of Variation, Deviation About the Mean, Average, Mean, Absolute Deviation

1. Introduction

Measures of dispersion or spread are one of the most important statistical data analysis tools. They are used in illustrating the distribution of datasets. The most important measures of dispersion are the measure of variation about the mean which shows the average spread of statistical data around the mean value [3]. The measure of variation about the mean have several applications such as use in theory of estimation to test the efficiency estimation techniques, used in statistical quality control to monitor quality controls process among other uses [3, 11].

Currently, there are three known measures of variation about the mean; Mean deviation which is the average absolute deviation about the mean, variance which is the average of squared deviations about the mean and standard deviation which is the square-root of average squared deviation about the mean [17]. Past studies have established that the existing measures of variation about the mean are not 100% efficient, this is due to various issues that arises during their use in estimating the average variation about the mean [2-3, 11-13, 16, 18, 20]. For example, mean deviation has been determined by past studies has to violate the algebraic number theory. Based on the algebraic number theory, given that the mean deviation is an average of absolute deviations, an absolute number on the field P such that $|\bullet| : p \to \Re^{>0}$

must satisfy the following conditions [5];

1. $|a| = 0 \ iff \ a = 0 \ \forall a \in p$
2. $|ab| = |a||b| \ \forall a, b \in p$
3. $|a+b| < |a| + |b| \ \forall a, b \in p$

The mean deviation about the mean is given by the function [17];

$$MAD = \frac{\sum_{i=1}^{n} |v_i - \bar{v}|}{n} \quad (1)$$

where $|v_i - \bar{v}|$ is the absolute deviation from the mean

This measure of deviation from the mean as an absolute number, violates the algebraic laws as illustrated by the third condition (3), hence the average deviation about the mean estimated by the measure are not accurate because;

$$\frac{\sum_{i=1}^{n} |v_i - \bar{v}|}{n} > \frac{\left|\sum_{i=1}^{n} (v_i - \bar{v})\right|}{n}$$

Therefore, the measure always gives bigger estimates than the actual deviation about the mean. However, the measure argues on the basis of the theory behind measurement of average deviation about the mean, which assumes that for measuring of deviation from the mean, the metric (the distance from the mean) is more important than the sign of the deviation. Mean deviation has also been determined by past studies not to allow further algebraic application because of the absolution, as a result the measure is not considered as efficient [6].

A second measure of variation about the mean is variance which is given by the function [1, 17];

$$\text{var} = \frac{\sum_{i=1}^{n} (v_i - \bar{v})^2}{n} \quad (2)$$

Where $(v_i - \bar{v})^2$ is the squared deviation from the mean

This measure of variation from the mean allows for further algebraic manipulations which is an improvement from the mean deviation, it also do not violate the algebraic number theory, however, the average of deviation about the mean given by the measure are squared, hence, are not of the same unit as the initial datasets (squared). This makes the results given by the formula to be inappropriate [1, 17].

The last measure of variation about the mean is standard deviation, which is an improvement on variance by giving results which are of the same units as initial datasets. The measure is usually estimated by [1, 3, 13, 16];

$$SD = \sqrt{\frac{\sum_{i=1}^{n} (v_i - \bar{v})^2}{n}} \quad (3)$$

Over the years, standard deviation has been the most widely used measure of variation about the mean, because it is a capable of further algebraic manipulation and it also solves the problem of variance by giving estimates which are of the same unit as the original datasets (square-root). However, past studies have determined that standard deviation is affected by outliers and skewed datasets, factors which makes this measure not to be efficient especially when dealing with datasets which have outliers and those that are skewed [2-3, 11-13, 16, 18, 20].

Given the shortcomings of the three existing measures of variation about the mean. The paper aimed at a modelling new measure of variation about the population mean known as geometric measure of variation, which is would overcome the weakness of the current measures of variation about the mean by not violating the algebraic laws, giving estimates which are of the same unit as the initial datasets, not affected by outliers and skewed datasets, and allows further algebraic manipulations to be carried on it.

2. Methods

An average measure of deviation is used to measure the average distant of each data point in a distribution from the mean of the dataset [17]. This help in determining the distribution of the dataset. The theory behind average deviation from the mean, believes that, in order to determine the distribution of data point from each other and from the mean, what is important is the absolute distance of each data point from the mean (metric) and not the direction of the deviation. Based on this theory, a small negative distance from the mean value is better than a large positive number, this is because datasets which compose of data points that are closer to the mean value are better than those that are composed of data points which are far away from the mean value [7]. This is because, what is important is important during statistical analysis is to obtain a measure of central tendency (mean) that is closer to all the data points in order to be a true representative of the entire dataset [16]. Therefore, the theory believes that it is statistically important to obtain estimates that are closer to all data Points [7]. As a result, when formulating a measure of variation from the mean, it is important to develop a measure that will capture the accurate magnitude of the deviation from the mean and not the direction of the variations. As a result, this study formulated a measure of variation from the mean which was aimed at estimating the average absolute deviations from the mean.

A good measure of variation from the mean should be one that obeys the algebraic laws, given that the study aimed at formulating a measure of variation about the mean which estimated the average absolute deviations about the mean, the measure needed to obey the three conditions that an absolute number must satisfy;

1. $|a| = 0 \ iff \ a = 0$
2. $|ab| = |a||b|$
3. $|a+b| < |a| + |b|$

In line with the algebraic laws of absolute numbers, the study formulated a measure which concentrated on the product of absolute deviations about the mean other than the sum so as not to violate the laws [5].

Lastly, past studies have determined that geometric averaging usually give averages that are not affected by outliers and also ones that do not assume symmetry of datasets [4, 14, 19, 21].

The theory behind geometric averaging believes that given a vector ϑ of identical data points ϑ_i such that $i = 1,2,3...n$, then the n[th] root of the product of all data points ϑ_i will yield the result ϑ which is an average representative of all the data points ϑ_i in the vector ϑ [14]. This can be illustrated mathematically as below;

Given that $\vartheta = [\vartheta_1, \vartheta_2, \vartheta_3, ...\vartheta_n]$ such that $\vartheta_1 = \vartheta_2 = \vartheta_3 = ... = \vartheta_n = \vartheta$ then the n^{th} root of the product of all data points is given as [14];

$$\text{Root} = \sqrt[n]{\prod \vartheta_i} = \vartheta \tag{4}$$

Which is a representative of all the data points because all data points are equivalent to ϑ [10].

Geometric averaging uses this theory to find the geometric mean for the data points which is given by as follows [8];

Consider a set of data points c_i such that $i = 1,2,3...n$ and $c_i > 0$ then the geometric mean for the data points is given by the formula;

$$GM = \sqrt[n]{\prod_{i=1}^{n} c_i} \tag{5}$$

Studies have found that compared to other means such as the harmonic mean and the arithmetic mean, the geometric mean always give results which are greater than the harmonic mean but also smaller than the arithmetic mean whenever all the data points are positive or greater than zero [4, 8, 15].

Therefore, the theory shows that geometric averaging can only be used when determining the average of positive numbers and the results is always between the arithmetic and the harmonic means. This is because if any data point is determined to be zero then its use in the formula will render the entire results impossible because 0^p is un defined. Also, if the data points have an odd p negative data points then the results $\left(\prod_{i=1}^{n} c_i\right)^{\frac{1}{n}}$ will yield a complex result which will not be part of the data points. As a result, during the formulation of the geometric measure of variation using the geometric averaging as an averaging technique, several alterations are carried out on the formula to ensure that the formulations do not give complex roots and also a rule on how to deal with zero deviations.

During the formulation of the geometric measure about the population mean, the study considered only the absolute deviations about the mean because during the study of deviations about the mean, what is important is the magnitude of the deviation and not the direction of the deviation. The formulation also ensured algebraic laws behind absolute numbers are obeyed during the formulations. Lastly, the formulation used geometric averaging which is not affected by outliers and skewed datasets in the averaging of the absolute deviations about the mean [4, 19, 21].

3. Results

3.1. Formulation of Geometric Measure of Variation for Un-weighted Datasets

For un-weighted datasets, consider a data vector $V = [v_i, v_2, v_3, ..., v_n]$ the geometric measure of variation about the mean for the data set can be derived as follows;

The arithmetic mean for the vector $\bar{v}$ is given by the formula;

$$\bar{v} = \frac{\sum_{i=1}^{n} v_i}{n} \tag{6}$$

The deviations d_i of each data point from the mean is given by;

$$d_i = v_i - \bar{v} \tag{7}$$

Hence the vector of deviations from the mean $d = [d_1, d_2, d_3, ..., d_n]$. Using the geometric averaging technique, the average deviation from the mean $\bar{d}$ can be obtained using the formula;

$$\bar{d} = \sqrt[n]{\prod_{i=1}^{n} d_i} \tag{8}$$

However, equation (8) is not applicable because given that mean is an equilibrium measure, the deviations from the mean tend to always be either, positive, negative or zero, as a result the product of the deviations can lead into a negative number which do not have real roots. In order to overcome this loophole, it would be better to use absolute deviations from the mean other than the actual deviations because according to the theory behind deviation, when assessing deviation, the magnitude is more important than the sign of the deviation. Now replacing the actual deviations with absolute deviations, the average deviation from the mean in (8) will be given by the formula;

$$\left|\bar{d}\right| = \sqrt[n]{\prod_{i=1}^{n} |d_i|} \tag{9}$$

The next problem that arises is when some deviations from the mean are zero which occur when the data points are zero, thus, even if one data point is zero the product of the deviations from the mean will be zero whose definite root do

not exist. In order to overcome these problem, the geometric deviation from the mean will average only those deviations from the mean which are none zero; Therefore, let P refers to all the none zero deviations from the mean, the average deviation from the mean using the geometric measure will be given by the formula;

$$\left|\bar{d}\right| = \sqrt[n]{\prod_{i=1}^{p}\left|d_i\right|} \tag{10}$$

Now given that when all the data points are the same then the average deviation from the mean would be zero. As a result, let G be the geometric measure of deviation from the mean. The average deviation from the mean for un-weighted datasets can be estimated as follows;

$$G = \begin{cases} \sqrt[n]{\prod_{i=1}^{P}\left|d_i\right|} & \forall d_i \neq 0 \\ 0 & \forall d_i = 0 \end{cases} \tag{11}$$

Using logarithms, equation (11), can be simplified to obtain equation (12) which is illustrated as below;

$$G = \begin{cases} \exp\left(\frac{1}{n}\sum_{i=1}^{P}\ln\left(\left|d_i\right|\right)\right) & \forall d_i \neq 0 \\ 0 & \forall d_i = 0 \end{cases} \tag{12}$$

The introduction of the logarithm assists in the elimination of infinite numbers that may rise as a result of the product of the deviations when the population size is too large, during the calculation of the geometric measure of variation from the mean.

3.2. Formulation of Geometric Measure of Variation for Weighted Datasets

Consider a vector of un weighted datasets $V = \left[v_1, v_2, v_3, ..., v_n\right]$ and a vector of weights $\varsigma = \left[\varsigma_1, \varsigma_2, \varsigma_3, ...\varsigma_n\right]$. When each element of the vector V is coefficient by the respective weight from the weight vector ς, a new vector of weighted dataset $\varsigma V = \left[\varsigma_1 v_1, \varsigma_2 v_2, \varsigma_3 v_3, ..., \varsigma_n v_n\right]$. The geometric measure of variation about the mean can be derived as follows;

First we begin by determining the mean for the weighted vector, this is calculated as follows;

$$\bar{V} = \frac{\sum_{i=1}^{n}\varsigma_i v_i}{\sum_{i=1}^{n}\varsigma_i} \tag{13}$$

The deviation from the mean will be given by;

$$d_i = v_i - \bar{v} \tag{14}$$

A vector of weighted deviations from the mean $\varsigma d = \left[\varsigma_1 d_1, \varsigma_2 d_2, \varsigma_3 d_3, ..., \varsigma_n d_n\right]$ can be derived by coefficient the weights for the respective data point and their respective deviations from the mean. The geometric average of variation from the mean for the weighted vector of deviations from the mean can be estimated using geometric averaging as follows;

$$\bar{d} = \sum_{i=1}^{n}\varsigma_i \sqrt{\prod_{i=1}^{n} d_i^{\varsigma_i}} \tag{15}$$

However, equation (15) just like (8) is not applicable because given that mean is an equilibrium measure, the deviations from the mean tend to always be either, positive, negative or zero, as a result the product of the deviations can lead into a negative number which do not have real roots. In order to overcome this loophole, it would be better to use absolute deviations from the mean other than the actual deviations because according to the theory behind deviation, when assessing deviation, the magnitude is more important than the sign of the deviation. Now replacing the actual deviations with absolute deviations, the average deviation from the mean in (15) will be given by the formula;

$$\left|\bar{d}\right| = \sum_{i=1}^{n}\varsigma_i \sqrt{\prod_{i=1}^{n}\left|d_i\right|^{\varsigma_i}} \tag{16}$$

The next problem that arises is when some deviations from the mean are zero which occur when the data points are zero, thus, even if one data point is zero the product of the deviations from the mean will be zero whose definite root do not exist. In order to overcome these problem, the geometric deviation from the mean will average only those deviations from the mean which are none zero; Therefore, let P refers to number of all the none zero deviations from the mean, the average deviation from the mean using the geometric measure will be given by the formula;

$$\left|\bar{d}\right| = \sum_{i=1}^{n}\varsigma_i \sqrt{\prod_{i=1}^{p}\left|d_i\right|^{\varsigma_i}} \tag{17}$$

Now given that when all the data points are the same then the average deviation from the mean would be zero. As a result, let G be the geometric measure of deviation from the mean. The average deviation from the mean for un-weighted datasets can be estimated as follows;

$$G = \begin{cases} \sum_{i=1}^{n}\varsigma_i \sqrt{\prod_{i=1}^{p}\left|d_i\right|^{\varsigma_i}} & \forall d_i \neq 0 \\ 0 & \forall d_i = 0 \end{cases} \tag{18}$$

Using logarithms, equation (18), can be simplified using logarithms to obtain equation (19) which is illustrated as below;

$$G=\begin{cases}\exp\left(\dfrac{1}{\sum_{i=1}^{n}\varsigma_i}\sum_{i=1}^{p}\varsigma_i\ln\left(\left|d_i\right|\right)\right) & \forall d_i\neq 0\\ 0 & \forall d_i=0\end{cases} \quad (19)$$

The introduction of the logarithm assists in the elimination of infinite numbers that may rise a result of the product of absolute deviations, this assist in the simplification of the geometric measure of variation from the mean.

3.3. Formulation of Geometric Measure of Variation for Probability Mass Functions

Based on equation (19), if all the deviations are none-zero then it can be determined that;

$$\ln\left(G\right)=\frac{1}{\sum_{i=1}^{n}\varsigma_i}\sum_{i=1}^{n}\varsigma_i\left|d_i\right| \quad (20)$$

Thus, $\ln\left(G\right)=E\left(\ln\left|d_i\right|\right)$ and that $\ln\left|d_i\right|$ is distributed with the same weights as v_i. Therefore, extending this relationship on probability mass functions. Assume that the variable v_i is discrete with probability mass function $\varsigma\left(v_i\right)$ for all $i=1,2,3,...,n$ and 0 otherwise. Assume that $\ln\left|d_i\right|$ which is equal to $\ln\left|v_i-\overline{v}\right|$ where $\overline{v}$ is the mean of the random variable v, is distributed in the same way as v_i with a probability mass function $\varsigma\left(v_i\right)$. Then the log of geometric deviation $\left(\ln\left(G\right)\right)$ can be shown based on equation (19) to be equivalent to;

$$\ln\left(G\right)=E\left(\ln\left|d\right|\right) \quad (21)$$

But $\ln\left|d_i\right|$ is distributed with probability mass function $\varsigma\left(v_i\right)$ then by definition;

$$\ln\left(G\right)=E\left(\ln\left|d_i\right|\right)=\sum_{i=1}^{n}\varsigma\left(v_i\right)\bullet\ln\left|d_i\right| \quad \forall d_i\neq 0 \quad (22)$$

Hence the geometric deviation for probability mass functions can be given as;

$$G=\begin{cases}\exp\left(\sum_{i=1}^{n}\varsigma\left(v_i\right)\bullet\ln\left|d_i\right|\right) & \forall d_i\neq 0\\ 0 & \forall d_i=0\end{cases} \quad (23)$$

3.4. Formulation of Geometric Measure of Variation for Probability Density Functions

Similarly, the relationship in equation (22) can be extended on continuous random variables. Assume that the variable v is continuous on the interval $a\le v\le b$ with probability density function $\varsigma\left(v\right)$. Assume that $\ln\left|d\right|$ which is equal to $\ln\left|v-\overline{v}\right|$ where $\overline{v}$ is the mean of the random variable v, is distributed in the same way as v with a probability density function $\varsigma\left(v\right)$.

$$\ln\left(G\right)=E\left(\ln\left|d\right|\right)=\int_{a}^{b}\varsigma\left(v\right)\bullet\ln\left|d\right|.dd \quad \forall d\neq 0 \quad (24)$$

Hence the geometric deviation for probability density functions can be given as;

$$G=\begin{cases}\exp\left(\int_{a}^{b}\varsigma\left(v\right)\bullet\ln\left|d\right|.dd\right) & \forall d\neq 0\\ 0 & \forall d=0\end{cases} \quad (25)$$

4. Conclusion

In conclusion, the study was able to formulate a geometric measure of variation about the population mean which can be used to estimate the average variation about the mean for un-weighted datasets, weighted datasets, probability mass functions and probability density functions. The results show that the formulations do obey the algebraic laws behind absolute numbers. The measure is also capable of further algebraic manipulations as it can be used further to estimate the average variation about the mean for weighted datasets, probability mass functions and probability density functions. Lastly, the geometric averaging ensures that the results obtained by the formulations are not affected by outliers and skewed datasets because it uses geometric averaging which has been determined by past studies not to be affected by skewed data sets and outliers [4, 19, 21].

References

[1] Ahn, S., & Fessler, J. A., (2003). Standard Errors of Mean, Variance, and Standard Deviation Estimators. EECS Department. The University of Michigan. U.S.A.

[2] Altman, D. G., & Bland, J. M. (2005). Standard deviations and standard errors. BMJ Volume 331.

[3] Bhardwaj, A., (2013). Comparative Study of Various Measures of Dispersion. Journal of Advances in Mathematics. Vol 1, No 1.

[4] Buckland, S. T., A. C. Studeny, A. E. Magurran, J. B. Illian, and S. E. Newson. (2011). The geometric mean of relative abundance indices: a biodiversity measure with a difference. Ecosphere 2 (9): 100. doi: 10.1890/ES11-00186.1.

[5] Clark, P. L., (2012). Number Theory: A Contemporary Introduction. Available at http://math.uga.edu/~pete/4400FULL.pdf

[6] Deshpande, S., Gogtay, N. J., Thatte, U. M., (2016). Measures of Central Tendency and Dispersion. Journal of the Association of Physicians of India. Vol. 64. July 2016.

[7] Grechuk, B., Molyboha, A., & Zabarankin M., (2011). Mean-Deviation Analysis in The Theory of Choice. Risk Analysis.

[8] Hu, S., (2010). Simple Mean, Weighted Mean, or Geometric Mean? Presented at the 2010 ISPA/SCEA Joint Annual Conference and Training Workshop.

[9] Kum, S., & Lim, Y., (2012). A Geometric Mean of Parameterized Arithmetic and Harmonic Means of Convex Functions. Hindawi Publishing Corporation. Volume 2012, Article ID 836804.

[10] Lawson, J. D., & Lim, Y., (2001). The Geometric Mean, Matrices, Metrics, and More. The American Mathematical Monthly.

[11] Lee, D., In, J., & Lee, S., (2015). Standard deviation and standard error of the mean. Korean journal of anesthesiology. 68. 220-3. 10.4097/kjae.2015.68.3.220.

[12] Leys, C., Klein, O., Bernard, P., & Licata, L., (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. Journal of Experimental Social Psychology 49 (2013) 764–766.

[13] Manikandan, S., (2016). Measures of dispersion. Journal of Pharmacology and Pharmacotherapeutics. October-December 2011. Vol 2. Issue 4.

[14] McAlister, D., (1879). The Law of Geometric Mean. The Royal Society is collaborating with JSTOR to digitize, preserve, and extend access to Proceedings of the Royal Society of London.

[15] Mindlin, D., (2011). On the Relationship between Arithmetic and Geometric Returns. Cdi Advisors Research. LLC.

[16] Mohini, P. B., &Prajakt, J. B., (2012). What to use to express the variability of data: Standard deviation or standard error of mean?. Perspectives in clinical research. July 2012.

[17] Raymondo, J., (2015). Measures of Variation from Statistical Analysis in the Behavioral Sciences. Kendall Hunt Publishing.

[18] Roberson, Q. M., Sturman, M. C., & Simons, T. L., (2007). Does the Measure of Dispersion Matter in Multilevel Research? A Comparison of the Relative Performance of Dispersion Indexes. Cornell University School of Hotel Administration. The Scholarly Commons.

[19] Roenfeldt, K., (2018). Better than average: Calculating Geometric Means Using SAS. Henry. M. Foundation for the Advancement of Military Medicine.

[20] Schuetter, J. (2007). Chapter 1. In J. Schuetter, measures of dispersation (pp. 45-54).

[21] Thenwall, M. (2018). The precision of the arithmetic mean, geometric mean and percentiles for citation data: An experimental simulation modelling approach. Statistical Cybermetrics Research Group, School of Mathematics and Computer Science, University of Wolverhampton, Wulfruna Street, Wolverhampton, UK.