Exploring the Geometric Mean of Grouped Data

Xiuqing Ji
Christopher Newport University

John Simon
Governors State University

We introduce a new method to estimate the geometric mean of grouped data by deriving an unbiased estimate based on uniformly distributed values within class intervals. In the current literature, traditional approaches often rely on class midpoints or geometric means of class endpoints, which may lead to biased results. Our proposed estimator is shown to perform more accurately, particularly when it is applied to financial data, such as the annual growth rates of the S&P 500 index. This study demonstrates the estimator's effectiveness through both analytical derivation and empirical testing, offering a valuable tool for analysis.

Keywords: geometric mean, grouped data, unbiased estimate, stock market data

INTRODUCTION

One of the earliest concepts that students learn in statistics is that of summary measures like the arithmetic mean, median, and the mode. Typically, computational formulas are presented first, and desired properties such as unbiasedness are addressed later, after an adequate preparation in probability theory. Likewise, computational formulas for raw data are introduced first, and later students are shown how to estimate the summary measures from grouped data.

Students get less exposure to summary measures like the geometric mean, and perhaps even less about its estimation from grouped data. Watier et. al. (2011) discusses the ways to address the meaning of mean, and the focus is only on the arithmetic mean. In specific contexts, other types of means like the geometric mean or harmonic mean are very useful, and students should be exposed to those (McNichol, 2018).

Geometric mean was brought into prominence perhaps by Francis Galton (1879) in the 19th century. It is a very useful summary statistic with respect to data about growth. Geometric mean of n positive real numbers is defined as the n'th root of their product. For example, suppose that a colony of 1000 microbes grew by 'multiples' of 1.2, 0.9, and 2.4 in three successive periods. By this we mean that at the end of the first period there were 1000 times 1.2 microbes (=1200), at the end of the second period there were 1200 times 0.9 (=1080) microbes, and at the end of the third period there were 1080 times 2.4 (=2592) microbes. The average (arithmetic mean) of the three multiples is 1.5, but this cannot be taken as the constant multiple each period. The geometric mean which is 1.3737 (= cube root of 1.2*0.9*2.4) is the appropriate summary measure, as we see here: 1000 * 1.5^3 = 3375, 1000 * 1.3737^3 = 2592.

For this reason, geometric mean finds extensive use in financial and economic studies (Costa, 2018) as well as population growth studies. Rapid growth in computing also calls for the use of geometric mean in describing the performance increase over the years (Fleming and Wallace, 1986).

In calculating the geometric mean, it may be easier to take the logarithms of the data values, take their arithmetic mean, and then take the anti-logarithm of that. Using the earlier example, the multiples are 1.2, 0.9, and 2.4; their logarithms (to base 10, although any base can be used) are 0.07918, -0.04576, and 0.38021; the arithmetic mean of those is 0.13788; and its anti-logarithm (i.e. raise 10 to that value) is the geometric mean 1.3737. It may be necessary to use this approach when computational accuracy is compromised by taking the product of many values.

Going from raw data to grouped data, a common approach (see for example https://www.hackmath.net/en/calculator/geometric-mean) in estimating the geometric mean is to follow the approach used in arithmetic mean and use the class midpoint as a representative value for the class. Another approach is to use the geometric mean of the class end points as the representative value for the class (see Szatrowski, 1946). In either case, it is not clear that these are unbiased estimates. In general, within the class interval, one may assume that the values are uniformly distributed (this is subject to debate). If this assumption is granted, we can derive an unbiased estimate for the geometric mean for the class interval and use that in estimating the geometric mean for the grouped data (for a discussion on bias, see White, 2019). These are carried out in the next two sections. In section 4, we will then apply the estimate to stock market data.

Geometric Mean of Uniformly Distributed Values in an Interval

Here we seek the expected value of the geometric mean of values uniformly distributed in an interval [L, U]. As mentioned earlier, the motivation for this comes from trying to find the geometric mean of grouped data, where we may assume that within a class interval, the values are uniformly distributed.

Let $X_1, X_2, \ldots, X_n$ be independent and uniformly distributed in the interval [L, U] (where $0 < L < U$, $n > 0$). We have the following result for the expected value of the geometric mean Gn:

$$E(Gn) = E\left(\sqrt[n]{\prod_{i=1}^{n} X_i}\right) = \int_L^U \int_L^U \ldots \int_L^U \frac{\sqrt[n]{\prod_{i=1}^{n} x_i}}{(U-L)^n}\, dx_1 dx_2 \ldots dx_n = \left(\frac{U^{1+\frac{1}{n}} - L^{1+\frac{1}{n}}}{(U-L)\left(1+\frac{1}{n}\right)}\right)^n \quad (1)$$

Using the L'Hospital's rule, we also have the limit of E(Gn) as n becomes large:

$$\lim_{n\to\infty} E(Gn) = \frac{U^{\left(\frac{U}{U-L}\right)}}{e\, L^{\left(\frac{L}{U-L}\right)}} \quad (2)$$

(where e is the base of natural logarithms).

For example, say that the values for a growth multiple are uniformly distributed in the interval [0.5, 2.0] (i.e. growth multiples range from half to double). Some select values of n and E(Gn) are given in Table 1.

TABLE 1
EXPECTED GEOMETRIC MEAN OF N UNIFORM RANDOM VARIABLES IN [0.5, 2.0]

n	1	2	3	5	10	50	infinity
E(Gn)	1.25	1.20988	1.19607	1.18489	1.17644	1.16965	1.16794

So, if three values are chosen at random in the interval [0.5, 2.0], then the expected value of their geometric mean is 1.19607, different from the arithmetic mean of the class end points which is 1.25, or the geometric mean of the class end points which is 1 (= square root of 0.5 * 2.0). Hence both 1.25 and 1 are

biased estimates. If a large number of values are chosen, then the expected value of the geometric mean is close to 1.16794.

A simulation using ten thousand iterations of generating three random variables uniformly distributed in [0.5, 2.0] and finding their geometric mean resulted in this 99% confidence interval for the geometric mean: (1.189, 1.202). This agrees with our findings.

However, it should be noted that the expected overall growth (the product of the growth multiples) over n values would be the arithmetic mean raised to n. This is because (assuming independence), E (X_1 * X_2 * … * X_n) is the n'th power of E(X_1). This can be confusing, but it is good to remember that E(f(X)) is often not equal to f(E(X)) for a given function f(x). Thus, if three growth multiples are chosen at random in the interval [0.5, 2], then the expected value of (X_1) (X_2) (X_3) would be 1.25^3.

Geometric Mean of Grouped Data

Say that we have grouped data, with classes (L_i, U_i) for i = 1 to k, class midpoints m_1, m_2, …, m_k, (where m_i = (L_i + U_i)/2) and with corresponding frequencies f_1, f_2, …, f_k (let $\sum_{i=1}^{k} f_i = n$). Then the arithmetic mean of the grouped data is estimated as $\frac{\sum_{i=1}^{k} m_i f_i}{n}$. Here the value m_i is being taken as a representative value for the i'th class interval. If we assume that the values in the class interval are uniformly distributed, then m_i is also the expected value of the arithmetic mean of the values in that class (i.e. $m_i = E\left(\frac{\sum X_j}{f_i}\right)$ where the summation is over the values in the class interval meaning 1 to f_i) and the estimate for arithmetic mean is unbiased.

Analogously, geometric mean for grouped data is often estimated as $\sqrt[n]{\prod_{i=1}^{k} (m_i)^{(f_i)}}$ (where m_i is again the class midpoint). If the values within a class are uniformly distributed, this is a biased estimate. Likewise, using the geometric mean of the class endpoints, we have $\sqrt[n]{\prod_{i=1}^{k} (h_i)^{(f_i)}}$ where $h_i = \sqrt{L_i U_i}$. This too is biased. This raises the question: what is the expected value of the geometric mean of f_i uniformly distributed values in a class interval? This was answered in the previous section and using that value instead of m_i or h_i should result in a better estimate.

Thus we propose a new estimate for the geometric mean estimate:

$$\sqrt[n]{\prod_{i=1}^{k} (g_i)^{(f_i)}}, \text{ where } g_i = \left(\frac{{U_i}^{1+\frac{1}{f_i}} - {L_i}^{1+\frac{1}{f_i}}}{(U_i - L_i)\left(1+\frac{1}{f_i}\right)} \right)^{f_i} \tag{3}$$

Here when f_i = is 0, we should take g_i to be 1 in order to avoid computational errors of division by zero. For ease of computation, we can find $\frac{\sum f_i \log(g_i)}{\sum f_i}$ and take its antilogarithm.

Using the Estimate on Stock Market Data

We will now apply the estimate proposed in the previous section to real data. The historical data of the US stock market index S&P 500 for the past 92 years is available online (for example, from S&P 500 (^GSPC) Historical Data, 2020). The closing value at the end of each year is noted in Table 2 (only a part of the data is shown to save space).

TABLE 2
ANNUAL CLOSING VALUE AND GROWTH OF S&P 500 INDEX

Date	Closing Value	Annual growth
31-Dec-2019	3,230.78	1.288781
31-Dec-2018	2,506.85	0.937627
29-Dec-2017	2,673.61	1.194200
30-Dec-2016	2,238.83	1.095350
.	.	.
.	.	.
.	.	.
31-Dec-1931	8.12	0.529335
31-Dec-1930	15.34	0.715152
31-Dec-1929	21.45	0.880903
31-Dec-1928	24.35	1.378822
30-Dec-1927	17.66	

The value at the end of the year 2019 was 3230.78, and at the end of the year in 2018 was 2506.85, hence the growth in 2019 was 3230.78 / 2506.85 = 1.288781. These growth ratios range in values from 0.5293 (in 1931) to 1.4502 (in 1954). A frequency distribution of the annual growth values for the 92 years is in Table 3.

TABLE 3
GROUPED DATA FOR ANNUAL GROWTH OF S&P 500 INDEX

L_i	U_i	f_i
0.5	0.75	5
0.75	1	26
1	1.25	42
1.25	1.5	19

Let us estimate the geometric mean for this data. Using the estimate $\sqrt[n]{\prod_{i=1}^{k}(m_i)^{(f_i)}}$, we get 1.05787 and using the estimate $\sqrt[n]{\prod_{i=1}^{k}(g_i)^{(f_i)}}$, we get 1.05532. Using the estimate $\sqrt[n]{\prod_{i=1}^{k}(h_i)^{(f_i)}}$, the result is 1.049739 (m_i, g_i, and h_i are defined in section 4). The real value? Since we have the original data, this is easy to compute: (3230.78 / 17.66) ^ (1/92) = 1.05457. We can see that in this case, the newly proposed estimate is closer to the real value. This need not always be the case, since the data may not be uniformly distributed within the class intervals.

DISCUSSION

Geometric mean is a useful summary statistic in situations that study growth, be it finance, biological populations, or computing performance. In comparison with the arithmetic mean, students do not get much

exposure to it, especially in the context of grouped data. Here we have explored its estimation in grouped data and derived an unbiased estimate for uniformly distributed values in an interval and its limiting value.

Instructors can engage students on the concept of bias in estimations, and for students with a calculus background, computing the expected value of the geometric mean of uniformly distributed values and its limit would be a good exercise. Simulation can be employed in exploring whether the arithmetic mean or the geometric mean of class end points are biased. Additionally, students can test these approaches on real data, such as the stock market returns.

REFERENCES

Costa, J. (2018, March 23). *Calculating Geometric Means (with online calculator).* Retrieved from https://buzzardsbay.org/special-topics/calculating-geometric-mean/

Fleming, P.J., & Wallace, J.J. (1986, March). How not to lie with statistics: The correct way to summarize benchmark results. *Commun. ACM29*, (3), 218–221.

Galton, F. (1879). The geometric mean, in vital and social statistics. *Proceedings of the Royal Society of London*, *29*(196–199), 365–367.

McNichol, D. (2018). *On Average, You're Using the Wrong Average: Geometric & Harmonic Means in Data Analysis.* Retrieved from https://medium.com/data-science/on-average-youre-using-the-wrong-average-geometric-harmonic-means-in-data-analysis-2a703e21ea0

S&P 500 (^GSPC) Historical Data. (2020). Retrieved from https://finance.yahoo.com/quote/%5EGSPC/history/?frequency=1mo&period1=-1325548800&period2=1577750400

Szatrowski, Z. (1946). Calculating the Geometric Mean from a Large Amount of Data. *Journal of the American Statistical Association*, *41*(234), 218–220.

Watier, N.N., Lamontagne, C., & Chartier, S. (2011). What does the mean mean? *Journal of Statistics Education*, *19*(2).

White, S.R., & Bonnett, L.J. (2019). Biased sampling activity: An investigation to promote discussion. *Teaching Statistics*, *41*(1), 8–13.