Paper 56-2018

Better than Average: Calculating Geometric Means Using SAS

Kimberly Roenfeldt
Henry M. Jackson Foundation for the Advancement of Military Medicine

ABSTRACT

Geometric means are a robust and precise way to visualize the central tendency of a data set, particularly when examining skewed data or comparing ratios. Measures of central tendency are predominantly presented as arithmetic means or medians that are relatively simple to calculate and interpret, but may be inaccurate in representing data that are not strictly normal. Geometric means represent the best of both worlds, providing estimates that take into account all the observations in a data set without being influenced by the extremes. They can be applied by SAS® programmers working in multiple industries including business, finance, health care, and research. Examples are varied and include examining compounded interest rates or returns on investments, assessing population changes in longitudinal data, or investigating lognormal data such lab assay results, biological concentrations, or decay rates. Fortunately, there are multiple methods in SAS that can be used to calculate geometric means including the GEOMEAN() function, the geomean keyword in PROC SURVEYMEANS, as well as manual data manipulations such as log transformation combined with PROC MEANS and exponentiation. This paper will explain the utility of geometric means and provide examples for using SAS to calculate geometric means and their measures of variability for your data.

INTRODUCTION

What is a geometric mean?

Geometric means are a type of “average”, or measure of central tendency in a distribution of data points, in the same group as the median, mode, or arithmetic mean. Whereas the arithmetic mean is calculated by summing a series of data points and then dividing that sum by the number of data points (Equation 1), the geometric mean *multiplies* a series of data points, and then uses the *n* number of data points to find the *n*^th^ *root* of that product (Equation 2). Mathematically, the geometric mean adds depth and stability to the mean.

$$\boldsymbol{Arithmetic\ Mean} = \left(\sum_{n=1}^{i} \boldsymbol{a_i}\right)/\boldsymbol{n} \; = \; {}^{\boldsymbol{1}}\!/_{\boldsymbol{n}}\,(\boldsymbol{a_1} + \; \boldsymbol{a_2} + \boldsymbol{a_3} + \cdots + \boldsymbol{a_n}) \qquad \textbf{(1)}$$

$$\boldsymbol{Geometric\ Mean} = \left(\prod_{i=1}^{n} \boldsymbol{a_i}\right)^{\boldsymbol{1}/\boldsymbol{n}} = \sqrt[n]{\boldsymbol{a_1 a_2 a_3} \dots\dots \boldsymbol{a_n}} \qquad \textbf{(2)}$$

We can easily visualize the geometric mean when applying it to its counterpart, the geometric series of numbers, where each number increases from the previous number according to the same proportion. The geometric mean will lie in the direct center of the values, whereas the arithmetic mean would have been “pulled” towards the higher values, and thus not truly represent the center of the data (Figure 1).

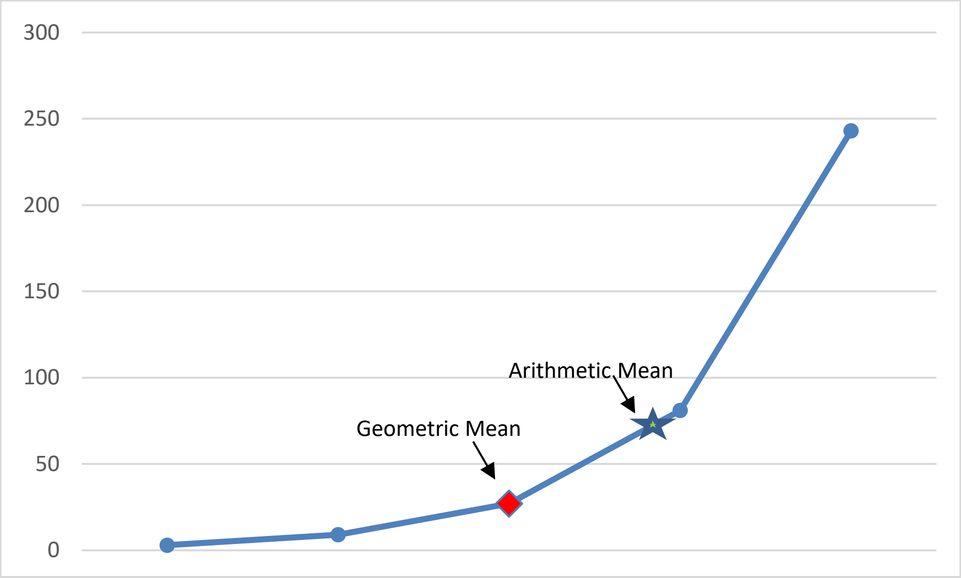


Figure 1: Geometric Series with Means Highlighted

$$Arithmetic\ mean = \frac{3+9+27+81+243}{5} = 72.6 \qquad \text{(See Equation 1)}$$

$$Geometric\ mean = \sqrt[5]{3*9*27*81*243} = 27 \qquad \text{(See Equation 2)}$$

When should I use the geometric mean instead of the arithmetic mean?

There are no hard rules for which mean you should use. Different types of averages can be used to express slightly different concepts: the center of the data, the values most often seen, and/or the typical "expected" values may or may not all be conveyed by the same measure. Data is rarely perfect, and you may need to look at several different types of averages to decide what works best for what you are trying to communicate with your data. But in general, geometric means are preferable when looking at skewed data, scaled data, or when averaging ratios. Some common applications include:

- Population growth
- Compounding interest
- Bioassays
- Radioactive decay
- Dose-response relationships
- Count data
- Time Series data
- Longitudinal data
- Repeated measures data
- Bioequivalence trials

If your data involve rate changes or changes over time, your data may be skewed. Often these data have a lognormal distribution, and the geometric mean describes the center of lognormal data perfectly. In addition to skew, you should also consider the size of your sample. When working with small samples, the sensitivity of the arithmetic mean can be problematic. The geometric mean might be a better central measure, as it will consider all of the data points, but without being subject to the same “pull” that can deteriorate the interpretation of the arithmetic mean (Figure 2):

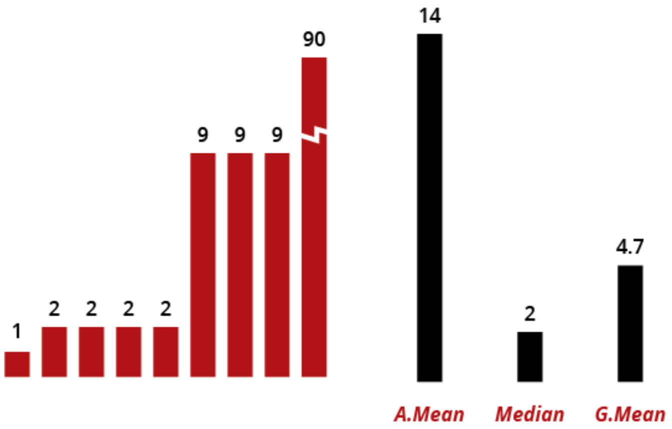


Figure 2: Comparison of Means for a Small Sample (McChesney, 2016)

In the above case, you should first confirm that 90 is a valid data point and not an error. Trimming outliers, so long as it is justifiable, is another way to produce more stable mean calculations.

Geometric means are also appropriate when summarizing ratios or percentages. This has many applications in medicine, and is considered the “gold standard” for calculating certain health measurements. In the financial industry, this concept is applied when constructing stock indexes and rates of return. The geometric mean is also employed in the art world, to choose aspect ratios film and video. The idea of comparing ratios is expanded when you look at scaled data: if you have data that have different attributes or scales, and you have normalized the results to be presented as ratios to reference values, the geometric mean is the correct mean to use.

CONSIDERATIONS

The calculation of the geometric mean requires that all values are non-zero and positive. So what should you do if you have data that do not meet this requirement? If you have values that equal zero, you have a few options:

1. Adjust your scale so that you add 1 to every number in the data set, and then subtract 1 from the resulting geometric mean.
2. Ignore zeros or missing data in your calculations.
3. Convert zeros to a very small number (often called “below the detection limit”) that is less than the next smallest number in the data set.

If you have negative numbers, you will need to convert those numbers to a positive value before calculating the geometric mean. You can then assign the resulting geometric mean a negative value. If your data set contains both positive and negative values, you will have to separate them and find the geometric means for each group, and you can then find the weighted average of their individual geometric means to find the total geometric mean for the full data set.

If none of these options appeals to you, you are not alone! There is controversy among statisticians about what is the best method for dealing with these values. You may want to calculate several types of averages and decide what makes the most sense for you and the results you are trying to report.

FINDING GEOMETRIC MEANS WITHIN AN OBSERVATION: GEOMEAN()/GEOMEANZ()

Finding the geometric mean of a series of data points is very easy using the GEOMEAN function with the following syntax:

GEOMEAN(*argument*<,*argument,…*>)

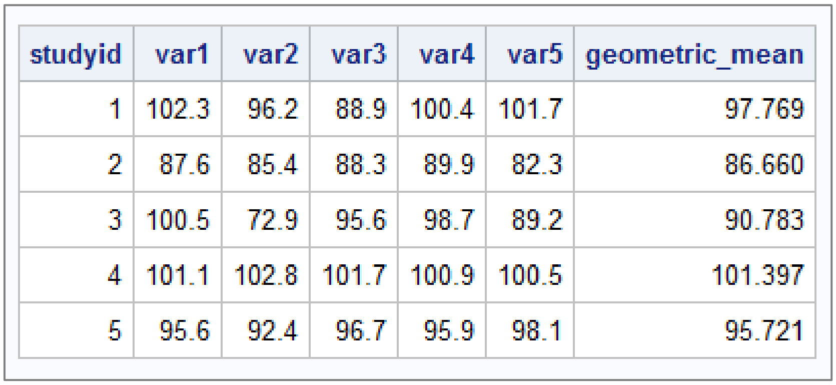
GEOMEAN will return the geometric mean of all non-missing values, and will “fuzz” any values that are close to zero: if one value is extremely small compared to the largest value, it is essentially treated as zero. If you do not want SAS to do fuzz values, then use the GEOMEANZ function, which has the same syntax. Consider the following example: you want to find the geometric mean of variables 1 through 5 for each participant in your data set. In the DATA step, you use the GEOMEAN function:

```
DATA my_data;
      input studyid var1 var2 var3 var4 var5;
         geometric_mean=geomean(of var1-var5); *Calculates geometric mean;
      datalines;
      1     102.3 96.2  88.9  100.4 101.7
      2     87.6  85.4  88.3  89.9  82.3
      3     100.5 72.9  95.6  98.7  89.2
      4     101.1 102.8 101.7 100.9 100.5
      5     95.6  92.4  96.7  95.9  98.1
      ;
run;
```

Note that you can use “OF” for a list of variables

Your results are displayed in the PRINT procedure:

```
PROC PRINT data=my_data;
   id studyid;
run;
```



studyid	var1	var2	var3	var4	var5	geometric_mean
1	102.3	96.2	88.9	100.4	101.7	97.769
2	87.6	85.4	88.3	89.9	82.3	86.660
3	100.5	72.9	95.6	98.7	89.2	90.783
4	101.1	102.8	101.7	100.9	100.5	101.397
5	95.6	92.4	96.7	95.9	98.1	95.721

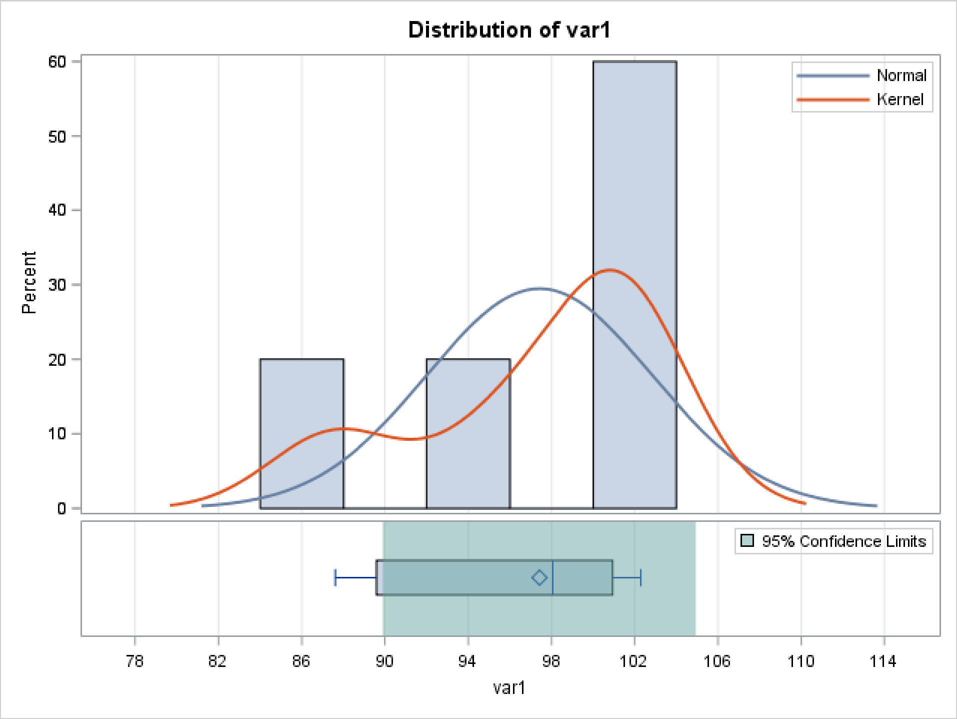
Output 1: Geometric Means by Row

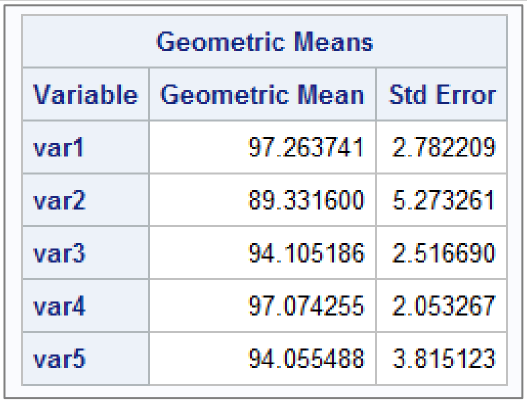
FINDING GEOMETRIC MEANS FOR A POPULATION

What if instead of finding the mean by each participant (i.e. by row), you wanted to find the geometric mean of each variable for the whole population (i.e. by column)? In this case, you would use the SURVEYMEANS procedure with a GEOMEAN statement:

```
PROC SURVEYMEANS data=my_data geomean;
   var var1 var2 var3 var4 var5;
run;
```

This procedure will return a lot of useful output including a histogram and boxplot of the distribution of each variable. The GEOMEAN statement additionally provides the geometric mean and its standard error for each specified variable in the VAR statement:





Geometric Means		
Variable	Geometric Mean	Std Error
var1	97.263741	2.782209
var2	89.331600	5.273261
var3	94.105186	2.516690
var4	97.074255	2.053267
var5	94.055488	3.815123

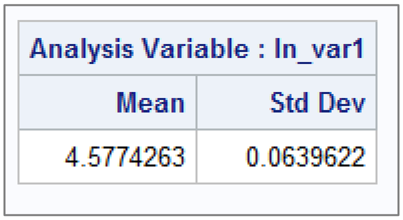
Output 2: Geometric Means by Column with Histogram and Boxplot (only var1 shown)

The standard error tells you how precise you can expect your geometric mean calculation to be. If you would rather have the geometric standard deviation as a way to envision the spread of the data, you need to do some data manipulations. Begin by finding the natural log of your variable in the DATA step using the LOG function (in this case, we will look only at variable 1):

```
DATA my_data2;
   set my_data;
      ln_var1=log(var1); *Calculates the natural log of variable 1;
run;
```

Next, the MEANS procedure will generate the arithmetic mean ("a_mean") and standard deviation ("a_stddev") of your log-transformed variable. Save these calculations into a new data set (here called "meansout") using the OUTPUT statement:

```
PROC MEANS data=my_data2 mean stddev; *Specifies output;
   var ln_var1;
   output out=meansout mean=a_mean stddev=a_stddev; *Creates new data set;
run;
```



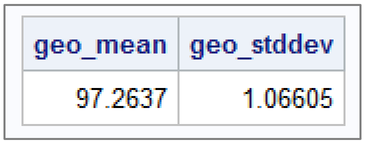
Analysis Variable : ln_var1	
Mean	Std Dev
4.5774263	0.0639622

Output 3: Arithmetic Mean and Standard Deviation of Log-Transformed Variable

You will then need to exponentiate the arithmetic mean and standard deviation to find the geometric mean and geometric standard deviation, using the EXP function in the DATA step:

```
DATA my_data3;
   set meansout;
      geo_mean=exp(a_mean); *Converts to geometric mean;
      geo_stddev=exp(a_stddev); *Converts to geometric standard deviation;
run;


PROC PRINT data=my_data3 noobs;
   var geo_mean geo_stddev;
run;
```



geo_mean	geo_stddev
97.2637	1.06605

Notice we have the same geometric mean for variable 1 that we found using PROC SURVEYMEANS, so we know we calculated the geometric mean correctly!

Output 4: Geometric Mean with Geometric Standard Deviation

Keep in mind that the geometric standard deviation is multiplicative: that is, the spread of your data is calculated by multiplying and dividing the geometric mean by the geometric standard deviation, rather than by adding to or subtracting from the geometric mean as you would do with the arithmetic standard deviation. So in the above example, one geometric standard deviation below the geometric mean is 91.2374 (Equation 4), and one geometric standard deviation above the geometric mean is 103.6880 (Equation 5):

$$Lower\ bound\ of\ Geometric\ Standard\ Deviation = {}^{Geometric\ Mean}\!/_{Geometric\ Standard\ Deviation} \quad \textbf{(4)}$$

$$= {}^{97.2637}\!/_{1.06605} = 91.2375$$

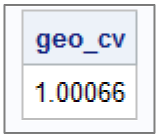
$$Upper\ bound\ of\ Geometric\ Standard\ Deviation = Geometric\ Mean \times Geometric\ Standard\ Deviation \quad \textbf{(5)}$$

$$= 91.2637 \times 1.06605 = 103.6880$$

If you are interested in comparing the variation of one data set with another, you use the geometric coefficient of variation. You just need to do another quick calculation in the data step, reducing the geometric standard deviation to the power of the reciprocal of the geometric mean:

```
DATA my_data4;
  set my_data3;
     geo_cv = geo_stddev**(1/geo_mean); *Calculates geometric CV;
run;

PROC PRINT data=my_data4 noobs;
   var geo_cv;
run;
```



geo_cv
1.00066

Output 5: Geometric Coefficient of Variation

CONCLUSION

If you are working with non-normal data, you should consider using the geometric mean as the measure of central tendency for your data. The geometric mean is a more robust and accurate way to find your average or expected value for data that is skewed, scaled, or proportional. The GEOMEAN statement within the DATA step, or the geomean option in PROC SURVEYMEANS, allow you to easily find geometric means using SAS.

REFERENCES

Alexander, N. (2012, June). Analysis of Parasite and Other Skewed Counts. *Trop Med Int Health, 17*(6), 684-693.

Crump, K. (1998, June). On summarizing group exposures in risk assessment: is an arithmetic mean or a geometric mean more appropriate? *Risk Analysis, 18*(3), 293-7.

Cudill, S. e. (2007). Geometric Mean Estimation from Pooled Samples. *Chemosphere*, 371-80.

Fleming, P. a. (1986, March). How Not to Lie with Statistics: The Correct Way to Summarize Benchmark Results. *Communications of the ACM, 29*(3), 218-221.

Franklin, D. (2016). "I Want the Mean, But not That One!". *PharmaSUG 2016 Conference Proceedings* (p. SP10). Denver, CO: Pharmaceutical SAS Users Group.

Glen, S. (2018). *Geometric Mean: Definition, Examples, Formula, Uses*. Retrieved from Statistics How To: http://www.statisticshowto.com/probability-and-statistics/

Habib, E. (2012, June). Geometric Mean for Negative and Zero Values. *International Journal of Recent Reseach and Applied Studies, 11*(3), 419-32.

Maddocks, J. (2018). *Growth and Decay - Geometric Growth and Decay*. Retrieved from Science Encyclopedia: http://science.jrank.org/pages/3154/Growth-Decay-Geometric-growth-decay.html

Matuszak, A. (2018). *Arithmetic, Harmonic, and Geometric Means with R.* Retrieved from The Economist at Large: http://economistatlarge.com/r-guide/arithmetic-harmonic-geometric-means-r

McChesney, J. (2016, December 15). *Why you should summarize data with the geometric mean.* Retrieved from Medium: https://medium.com/@JLMC/understanding-three-simple-statistics-for-data-visualizations-2619dbb3677a

Olivier, J. e. (2008, April). "The logarithmic transformation and the geometric mean in reporting experimental IgE results: what are they and why to use them? *Annals of Allergy, Asthma, and Immunology, 100*(4), 625-6.

OPS Systems. (2011, November 07). *Handling Zeros in Geometric Mean Calculation*. Retrieved from Water & Wastes Digest: https://www.wwdmag.com/channel/casestudies/handling-zeros-geometric-mean-calculation

Rasheed, A. a. (2015). Use of Geometric Mean in Bioequivalence Trials. *International Journal of Statistics in Medical Research, 4*, 114-120.

SAS Documentation. (2016). *SAS 9.4 Functions and CALL Routines, 5th Edition.* Cary, NC: SAS Institute.

Sawant, S. a. (2011). FAQ: Issues with Efficacy Analysis of Clinical Trial Data Using SAS. *PharmaSUG 2011* (p. PO08). Nashville, TN: Pharmaceutical SAS Users Group.

Sharma, J. (2014). *Business Statistics, Second Edition.* Jangpura, New Delhi: Vikas Publishing House Pvt Ltd.

Stafford Johnson, R. (2014). *Equity Markets and Portfolio Analysis.* Hoboken, NJ: John Wiley & Sons, Inc.

Walker, G. a. (2010). *Common Statistical Methods for Clinical Research with SAS Examples; 3rd Edition.* Cary, NC: SAS Institute, Inc.

ACKNOWLEDGMENTS

The author wishes to thank Rudolph Rull and Isabel Jacobson for their support, advice, and encouragement!

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Kimberly Roenfeldt
Henry M. Jackson Foundation for the Advancement of Military Medicine
(619) 767-4584
kimroenfeldt@gmail.com