

“I Want the Mean, But not That One!”

David Franklin, Quintiles Real Late Phase Research, Cambridge, MA

ABSTRACT

The ‘Mean’, as most SAS programmers know it, is the Arithmetic Mean. However, there are situations where it may be necessary to calculate different ‘means’. This paper first looks at different methods that are widely used from a programmer’s perspective, starting with the humble Arithmetic Mean, then proceeding to the other Pythagorean Means, known as the Geometric Mean and Harmonic Mean, before ending with a quick look at the Interquartile Mean and its related Truncated Mean. During the journey there will be examples of data and code given to demonstrate how each method is done and output.

INTRODUCTION

“I want the mean, but not that one!” Some of us have heard that phrase, others have not.

As a programmer we use procedures such as PROC MEANS or PROC UNIVARIATE, and occasionally write datastep code, to calculate the mean. But these ways of calculating the mean calculate the Arithmetic Mean. However, other methods exist.

This paper looks at first that the Arithmetic Mean, then proceeds to the Geometric Mean and Harmonic Mean, before ending with a quick look at the Interquartile Mean and its related Truncated Mean.

THE METHODS

We have all heard of the humble ‘mean’ or ‘average’, but what exactly is it?

When we hear of ‘mean’ or ‘average’ we are actually often thinking of the Arithmetic Mean (AM) which is defined as the sum of the non-missing values in a set divided by the number of non-missing values in the same set. The formula is often seen in textbooks as

$$\frac{1}{n} \sum_{i=1}^n a_i$$

The Arithmetic Mean is the most commonly used and readily understood measure of central tendency, and is best when the data is not skewed (having no extreme outlier values) and the individual data points are not dependent on each other.

But did you know there are other ‘means’ out there?

The first is the Geometric Mean (GM) which uses the product of the values as opposed to the arithmetic mean which uses their sum. This formula is often seen as

$$\left(\prod_{i=1}^n a_i \right)^{1/n}$$

The Geometric Mean should be used whenever the data are interrelated, usually ratios or percentages, and usually involve a time component.

As an example, let’s look at the growth rate of bacteria over a four hour period (as a ratio):

Time 0-60 minutes = 1.2 (100 start, 120 end)

Time >60-120 minutes = 1.4 (120 start, 168 end)

Time >120-180 minutes = 1.1 (168 start, 185 end)

Time >180-240 minutes = 0.4 (185 start, 74 end)

$$GM = \sqrt[4]{1.2 \times 1.4 \times 1.1 \times 0.4} = \sqrt[4]{0.7392} = 0.927$$

This is different to the arithmetic mean which would be

$$AM = \frac{1.2+1.4+1.1+0.4}{4} = \frac{4.1}{4} = 1.025$$

One shows an average decrease of 7.3% uniformly across each period while the other shows an increase of 2.5%. Which is more correct? $100 \times (0.927)^4 = 74$, the Geometric Mean.

The main reason for using the Geometric Mean over the Arithmetic Mean is that in the example the starting value of the second period is dependent on the first result.

The second mean we will look at is the harmonic mean, sometimes called the subcontrary mean) which uses reciprocals, with the formula often seen as

$$\frac{n}{\sum_{i=1}^n \frac{1}{x_i}}$$

This statistic is useful in situations involving data where the majority of the values are distributed uniformly but there are a few outliers with significantly higher or lower values. The Harmonic Mean gives less significance to high-value outliers, providing a truer picture of the average.

Lets look at some data:

$$HM = \frac{10}{\frac{1}{5} + \frac{1}{2} + \frac{1}{7} + \frac{1}{1} + \frac{1}{2} + \frac{1}{9} + \frac{1}{4} + \frac{1}{5} + \frac{1}{21} + \frac{1}{7}} = \frac{10}{3.094} = 3.23$$

is different to the arithmetic mean which would be

$$AM = \frac{5+2+7+1+2+9+4+5+21+7}{10} = \frac{63}{10} = 6.3$$

In general the harmonic average is less biased due to a small number of outliers.

The Arithmetic Mean, Geometric Mean and Harmonic Mean methods are considered part of what is often called the Pythagorean means.

Another method that is useful when dealing with outlier data, as shown in the data above, is to calculate the Interquartile Mean (IM), calculated using the formula

$$\frac{2}{n} \sum_{i=\frac{n}{4}+1}^{\frac{3n}{4}} x_i$$

In this method the first and fourth quartiles are removed and the arithmetic mean is done on the records in the second and third quartiles. One useful reason for removing the first and fourth quartiles is that and outliers are removed.

Looking at the data above and putting into order we get

1 2 2 4 5 5 7 7 9 21

Using the formula above, the Interquartile Mean is

$$IM = \frac{2}{10} \sum_{(10/4)+1}^{3*10/4} x_i = \frac{2}{10} \sum_4^7 x_i = \frac{2}{10} (4 + 5 + 5 + 7) = 4.2$$

A variation of the Interquartile Mean which incorporates the simplicity of the Arithmetic Mean is the Truncated Mean (TM) where numbers are excluded which are outside specified percentiles. Using this method we get bounds by removing particular observations. If we were to remove the top and bottom observations using a 10% trimmed mean, we would remove 10% of the observations on both sides then calculate the mean from the resulting data. Using our example above and a 10% truncated mean, taking 10% of the observations off each side result in the 1 and 21 are excluded, so the Truncated Mean can be computed

$$TM = \frac{2+2+4+5+5+7+7+9}{8} = \frac{41}{8} = 5.125$$

As said previously, the Harmonic Mean, Interquartile Mean and Truncated Mean are useful if there are outliers which strongly influence the result.

Which is correct, well that depends on the data and the statistician – programmers, follow any guidance from the statistician.

What does SAS do? This is an interesting question.

Using SAS procedures the only mean that are available is the Arithmetic Mean (a number of procedures) and the Geometric Mean via the SURVEYMEANS procedure.

Using the SAS functions in SAS 9.4, there is the MEAN, GEOMEAN and HARMEAN functions, but the problem with using these is that you have to have the data transposed to a row in order to use these – while our examples have this form, real world we do not have this option. Now we have to write some code using a procedure and a datastep.

To help understand the code with the examples above, I am going to use the same data in each example.

First we will look at the Arithmetic Mean and the data we used for the Harmonic Mean, to get both of these means:

```
data _dat;
  input x @@;
datalines;
5 2 7 1 2 9 4 5 21 7
;
run;
data _null_;
  set _dat;
  length numx sumx rcplx 8; retain numx sumx rcplx;
  set _dat end=eof;
  if _n_=1 then do;
    numx=0; sumx=0; rcplx=0;
  end;
  if ^missing(x) then do;
    numx+1; sumx+x; rcplx=rcplx+(1/x);
  end;
  if eof then do;
    AM=sumx/numx; HM=numx/rcplx;
    put NUMX= / SUMX= / AM= / HM=;
  end;
run;
```

which outputs the following values to the SAS LOG

```
numx=10
sumx=63
AM=6.3
HM=3.2315978456
```

This is what we got by hand.

Because the data used in the Geometric Mean is different than that for the other two, it is impossible to use the same data, so we go back to our original example data

```
data _dat;
  input x @@;
datalines;
1.2 1.4 1.1 0.4
;
run;
data _null_;
  length numx prodx 8; retain numx prodx;
  set _dat end=eof;
  if _n_=1 then do; numx=0; prodx=1; end;
  if ^missing(x) then do; numx+1; prodx=prodx*x;
  end;
  if eof then do;
    GM=prodx**(1/numx); put NUMX= / GM= ;
  end;
run;
```

which outputs the following values to the SAS LOG

```
numx=4
GM=0.9272364372
```

This is the same result we got by hand.

For the the Interquartile Mean, we will go back to our original data that we used for the Harmonic Mean, and get the following output:

```
data _dat;
  input x @@;
datalines;
5 2 7 1 2 9 4 5 21 7
;
run;
proc sort data=_dat;
  by x;
run;
data _null_;
  set _dat nobs=nobs end=eof;
  if ((nobs/4)+1)<=_n_<=(3*nobs/4) then sumx+x;
  if eof then do;
    IM=(2/nobs)*sumx;
    put nobs= / sumx= / IM=;
  end;
run;
```

which outputs the following values to the SAS LOG

```
nobs=10
sumx=21
IM=4.2
```

This is the same result we got by hand.

For the the Truncated Mean, we go back to the previous data and do a Truncated Mean of 10%, which means we drop 10% of the observations each side, and do

```
data _dat;
  input x @@;
datalines;
1 2 2 4 5 5 7 7 9 21
;
run;
proc sort data=_dat;
  by x;
run;
data _null_;
  set _dat nobs=nobs end=eof;
  if ((nobs*0.1)+1)<=_n<=(nobs*0.9) then do;
    numx+1;
    sumx+x;
    put x= numx= sumx= _n=;
  end;
  if eof then do;
    TM=sumx/numx;
    put numx= / sumx= / TM=;
  end;
run;
;
```

to get in the SAS LOG

```
numx=8
sumx=41
TM=5.125
```

which we got by hand.

CONCLUSION

Key to using statistical software is knowing what it can and cannot do. It is very interesting that SAS itself does not calculate the Geometric Mean or Harmonic Mean using the MEANS or UNIVARIATE procedure, nor have an option for either the Interquartile Mean or Truncated Mean. Maybe in the next SAS Software Ballot we can ask SAS to apply this option?

CONTACT INFORMATION

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

Name: David Franklin
Enterprise: Quintiles Real Late Phase Research
E-mail: David.Franklin@Quintiles.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.