Issues in Metric Selection and the TPC-D Single Stream Power

Alain Crolotte

AT&T Global Information Solutions

The purpose of this paper is to examine some of the issues faced by benchmark designers in situations where potentially very large and very small numbers need to be aggregated into a single number.We use, as an example, the case of TPC-D and examine the solution retained as described in the TPC-D Draft 6.0 [1]. The paper focuses on basic issues which are at the core of the metric selection problem and how choices on the basic issues naturally lead to choices on metric alternatives. To simplify the discussion, a mathematical appendix (Appendix A) has been provided thus freeing up the text from mathematics as much as possible.

The Problem In a situation where test results for a single vendor can be distributed over a very wide range, how can one aggregate all the results into a single figure of merit so that (1) the underlying business model is represented in the metric and (2), meaningful vendor to vendor comparisons can be performed? Also, we assume that "small is good" i.e., a low metric equates to a good score. This is the case for the 19 TPC-D queries [1] which represent a large sample of realistic business questions in a decision-support environment. For a given system and a given database size, query execution times can vary over a wide range, e.g. some queries could take a few seconds while others could take several minutes or even hours. Small individual observations should, of course, equate to a good score. Averaging, in some fashion, the observations, i.e. finding a characteristic of central tendency, and taking the inverse will provide such a score.

The simplest characteristic of central tendency is the simple mean which is equal to the sum of all observations (query times for TPC-D) divided by the number of observations. The main advantage of using the simple mean is its "physical" significance. For TPC-D its inverse is the average number of queries processed per unit of time. One of the well- known drawbacks of the arithmetic average is its sensitivity to large observations which are out of scope. For instance, if all observations are within the 1 to 10 range an abnormal observation of 1000 could dominate the arithmetic average so much that the resulting value could be meaningless. In this kind of a situation it is customary to resort to the geometric average. However, the geometric average is not a universal panacea as we see in the sequel.

For a given set of n observations (e.g. TPC-D query times) x_i 's, the simple mean $\bar{x}$ is given by

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}$$

and the geometric average g is given by the formula g = $\sqrt[n]{x_1 x_2 ... x_n}$ (the query power metric is computed as the inverse of g.) The formula defining g looks simple but it is somewhat confusing when one tries to understand its physical meaning. By taking the logarithm and bearing in mind its elementary properties (i.e and), the formula for g can be rewritten as

$$\log g = \frac{\log x_1 + \log x_2 + ... + \log x_n}{n}$$

Viewing the geometric average in this fashion provides a more intuitive approach to understanding a metric based on a geometric average. This is a result of the fact that adding is "easier" than multiplying, the very reason why logarithms were invented. With this formula we see that the geometric average can be viewed as an average also! Remember that the simple mean is very sensitive to large out of scope values. The same applies to the geometric average except that the sensitivity is to very small values of x resulting in very large (negative) values of logx.

More Averages In summary, so far, we have the arithmetic average which is too sensitive to large values and the geometric average which is too sensitive to small values. This makes the geometric average sensitive to "benchmark specials". A vendor having found a tricky way to make one observation extremely small would reap enormous benefits. As we saw earlier, the problem with the geometric average comes from the "discontinuity" of the logarithm at the origin. So long as the observations are away from zero the geometric average is well-behaved. To try staying away from zero let us add a small quantity f to all the x_i 's and to g which becomes say g_f defined by

$$\log (g_f + f) = \frac{\log (x_1+f) + \log (x_2+f) + \ldots + \log (x_n+f)}{n}$$

In this formula f is a fixed quantity independent of the observations. Although this formula looks identical to the "correction formula" portrayed in paragraph 5.4.1.2 of the TPC-D Draft 6.0 [1], it is different in the sense that, in the Draft, the quantity f is not fixed but a function of the observations. Since there are major drawbacks in using a value of f which is not fixed (see Appendix B) we confine ourselves to fixed values of f for the purpose of this analysis.

We have "coined" the term f-displaced geometric average for the quantity defined in the above formula. The purpose of the displacement is to guard against benchmark specials. But then, what to chose for f? We have retained two candidates, f=1/1000 and f=1. The first choice is reminiscent of the max/min ratio in the "correction formula" and there is a good reason for the second choice explained in Appendix A. Then, there is the half-way average called this way because it is a compromise, half-way between the simple mean and geometric average (again see Appendix A). The half-way average s is defined by the equation:

$$\sqrt{s} = \frac{\sqrt{x_1} + \sqrt{x_2} + \ldots + \sqrt{x_n}}{n}$$

Dimensions of Value The question we can ask ourselves next is: What are the properties which make a metric "good"? Next, we have defined five dimensions of value so that we can assess the above defined averages or any other metric, in terms of these dimensions of value. These are: (1) Ease To Explain referring to the amount of difficulty one encounters when explaining the metric to non-mathematically oriented users - it is related to how intuitive the measure is (everybody understands the simple mean). (2) Meaning refers to the ability to translate the measure into something usable directly while doing data base processing (e.g. a transaction rate). (3) Non-hypersensitivity To Extreme Values which refers to the propensity of a metric to be overwhelmed by certain values out of range. (4) Scalability which refers to the property of a metric to be scaled by the same amount as the individual observations (e.g. if all values are divided by 2 then the metric is divided by 2). (5) Balance which refers to the property of a metric to favor a relative decrease in a large value over the same relative decrease in a small value.

Ease To Explain and Meaning These two dimensions are very closely related. Of all the measures considered only the simple mean has meaning since it easily translates into a number of data base transactions a user could expect to perform in a unit of time. Of course the actual client workload may not be represented adequately but this is a general problem and customers can exercise responsibility and make sure they understand their workload and how it would affect their throughput. It will be difficult to explain any metric other than the simple mean. In the case of TPC-D, although all averages examined have the dimension of a query time they don't have meaning because they cannot be translated into a number of transactions one can perform. They can, however, be effective for summary and comparison purposes.

Hypersensitivity to Extreme Values The half-way average is not as sensitive as the simple mean to large values. For example, taking {1,2,...10} and bringing 1 to 1000, the simple mean is multiplied by 20 but the half-way average is multiplied by 6 while the geometric average is multiplied by 2 only. For small values the geometric and the displaced geometric averages behave similarly.

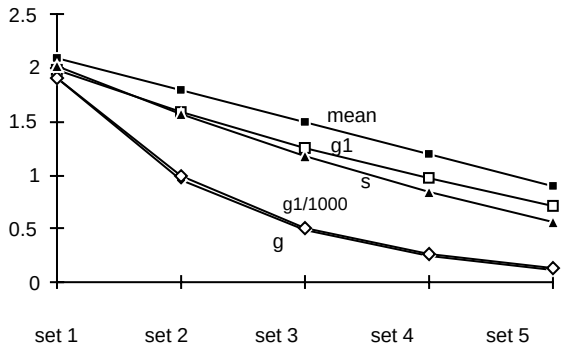
Scalability This property is important. All averages examined so far, except the f-displaced geometric average, scale with the observations; i.e. if all observations are multiplied (or divided) by a factor K then the average is also multiplied (or divided) by the same factor K. Non scalability is the major drawback of the 1-displaced geometric average which looks so good

otherwise, especially considering the fact that if all observations are small then the 1-displaced geometric average is close to the simple mean.

Balance What we have called here a balanced measure is one which rewards "working" on large observations. Of all measures considered only the unbalanced measure is the geometric mean. In other words, with the geometric mean it does not pay to "work" on the system to bring the larger observations down because a relative drop of say 10% in a large observation will equate in the same drop for the geometric average as would a drop of 10% in a larger observation. This was what the TPC-D subcommittee wanted to accomplish and that is why the TPC-D metric is based on a geometric aver- age. Table 1 and the graph under it show a comparison between the simple mean, the new measures and the geometric average in order to illustrate some of the points just mentioned.

TABLE 1. Comparison Between Arithmetic Average, Geometric Average, f- Displaced Geometric Average and Half-Average For Five Sets of Hypothetical Data

	set 1	set 2	set 3	set 4	set 5
x_1	1	1	1	1	1
x_2	1	1	1	1	1
x_3	1	1	1	1	1
x_4	2	2	2	2	2
x_5	2	2	2	2	2
x_6	2	2	2	2	2
x_7	3	3	3	3	0.003
x_8	3	3	3	0.003	0.003
x_9	3	3	0.003	0.003	0.003
x_{10}	3	0.003	0.003	0.003	0.003
mean	2.1	1.8003	1.5006	1.2009	0.9012
g_1	1.9804	1.5953	1.2600	0.96807	0.7138
g	1.9105	0.9575	0.4799	0.2405	0.1205
$g_{1/1000}$	1.9107	0.9850	0.5076	0.2613	0.1343
s	2.0081	1.5609	1.1699	0.8352	0.5568



In set 1 all observations are of the same order of magnitude and the values of the five statistics are very similar. In set 2 an observation has been brought down artificially through the use of a "benchmark special". The net effect on the geometric average is a division by 2 (corresponding to a doubling of the power which is the inverse) while the 1-displaced average or the half-way average behave similarly to the simple mean and decrease moderately. The 1/1000-displaced geometric average behaves also identically to the geometric average. As the number of

observations which are brought down increases, the high sensitivity of the geometric average is exemplified resulting in doubling the power every time while the increase using the 1-displaced geometric average or the half-way average is moderate, reflecting more of an "additive" effect.

Where do we go from here? Table 2 portrays the summary of the analysis which took place above. Based on the desired effect, what is important and how this translates into the above dimensions of value, a choice can be made. The usefulness of what is dis- cussed here is the focus on the real issues (the dimensions of value). In the case of TPC- D, for example, a major consideration was the unbalance (N in the balance column). The resulting choice was the geometric average and the advantages and drawbacks which come with it. But, instead of debating choices on metric types, people can debate the real issues and the relative importance of these issues. Table 2 can then assist in selecting a metric once the real issues and the choices on these issues are clear. My per- sonal choice in a decision support environment is the half-way average. On one hand I am afraid of benchmark specials because they have a tendency to cast a doubt on the entire benchmark process so I am afraid of the geometric average in spite of the very good arguments for it. On the other hand, I like the arithmetic average because it has meaning but it is overwhelmed by large values. The half-way average is right in between, it is not sensitive to extreme values and it scales. Therefore it is the right choice.

TABLE 2. Comparison Summary for Considered Measures

Measure	Ease to Explain	Meaning	Non-Hypersensitivity to Extreme Values	Scalability	Balance
current [1]	N	N	N	N	N
simple mean	Y	Y	N	Y	Y
XXX	N	N	Y	N	Y
XXX	N	N	N	Y	N
XXX	N	N	N	N	Y
s	Y/N	N	Y	Y	Y

APPENDIX A

The phi-average. Given a set of n observations x_1 ,..., x_n , and their associated weights or frequencies f_1 , .. ,f_n one can define a gamut of averages. A very broad range of such averages fall under the general category of phi-averages defined as in [2]: given a monotonic function ϕ the phi-average M_ϕ is given by the formula

$$\phi(M_\phi) = \sum_{1}^{n} f_i \, \phi(x_i)$$

The r-average (Also known as Power Average of Order r).

The r-average m_r is a special case of ϕ: x--> x^r and is given by

$$m_r = \left(\sum_i f_i \, x_i^r \right)^{1/r}$$

Important subcases are r = 1 (the arithmetic average often noted as $\bar{x}$), r = -1 (the harmonic average), and r = 0 which is a limit case yielding the geometric average (g), and r = 1/2 (the half-way average s). We now proceed to show that h < $\bar{x}$ < g. First we show that $\bar{x}$ < g by noticing that x --> log x is concave and that, therefore

$$\log \sum_i f_i \, x_i \;=\; \sum_i f_i \log x_i$$

The above equation merely expresses log g = log $\bar{x}$ which implies g = $\bar{x}$. By rewriting the above equation with y_i = we obtain yet another way of writing log g = log h where g and h are the geometric and harmonic averages of the 1 / x_i 's. The relationship between the harmonic average and the geometric average can be further exploited to show that the r-average is an increasing function of r. Taking the derivative with respect to r in the equation defining m_r yields

$$\frac{dm_r}{dr} = \frac{m_r}{r^2}\left(-\log \sum f_i x_i + \frac{\sum f_i \log x_i}{\sum f_i x_i}\right)$$

where $y_i = x_i^r$.

Setting $g_i = f_i y_i / \sum f_i x_i$ we can further reduce the above equation to

$$\frac{dm_r}{dr} = \frac{m_r}{r^2}\left(-\log \sum f_i x_i + \sum g_i \log y_i\right)$$

Calling H the harmonic average of the 1 / y_i 's with weights g_i the above equation can be rewritten as

$$\frac{dm_r}{dr} = \frac{m_r}{r^2}\left(\log H - \sum g_i \log \frac{1}{y_i}\right)$$

which is positive since the second term inside the parentheses in the equation above is log G where G is the geometric average of the 1 / y_i 's with weights g_i . Therefore, the derivative of m_r with respect to r is positive and therefore m_r is an increasing function of r.

The geometric average as the 0-average.

When r --> 0 we can write $a^r = e^{r \log a} = 1 + r \log a + o(r)$ and thus:

$$m_r^r = 1 + r \log m + o(r) = \sum_i f_i (x_i)^r = \sum_i f_i + r \sum_i f_i \log x_i + o(r)$$

which simplifies into

$$r \log m_r = r \sum_i f_i \log x_i + o(r)$$

finally yielding

$$\log m_0 = \sum_i f_i \log x_i$$

which is the definition of the geometric average, the phi-average for ϕ: x --> logx. As a result, the half-way average corresponding to r = 1/2 is half-way between the geometric average (r = 0) and the simple average (r = 1).

The1-displaced average.

Consider a phi-average closely related to the geometric average namely the a-displaced geometric average defined by

$$\log (a + g_a) = \sum_i f_i \log (a + x_i)$$

where a is positive. Considering the 1-displaced geometric average from the point of view of its mathematical properties, we have already noticed that it falls into the category of phi-averages corresponding to the function x--> log (a + x); but, in this family of functions (a being the parameter), the 1-displaced geometric average (corresponding to a = 1) plays a role of anchor because it is the only one for which the value of the function is zero when the variable is equal to zero. This is why it was retained as a candidate.

Relationship between arithmetic, geometric and f-displaced geometric averages

The geometric average can be denoted as g_0 for consistency. It could also be denoted as m_0 since it is also the 0-average, and for this reason, since we know that the r-average is an increasing function of r, we have $g_0 = \bar{x}$. We are now showing that g_a is between g_0 and $\bar{x}$. First, notice that x --> log (a + x) is concave and that therefore

$$\log \sum_i f_i (a + x_i) = \sum_i f_i \log (a + x_i)$$

Hence, $\log (a + \bar{x}) = \log (a + g_a)$ and, since x --> log (a + x) is also monotonic increasing, then $\bar{x} = g_a$ and this establishes the first part of the inequality.

To establish the second part of the inequality consider the a-displaced geometric average g_a defined by

$$\log (a + g_a) = \sum_i f_i \log (a + x_i)$$

$a + g_a$ can also be interpreted as G(y), the geometric average of the y_i 's defined by $y_i = a + x_i$. Taking the derivative with respect to a in the equation defining g_a yields

$$\frac{1 + \frac{d}{da} g_a}{a + g_a} = \sum_i f_i \frac{1}{a + x_i}$$

which can be rewritten as

$$\frac{1}{G(y)} \frac{d}{da} g_a = \sum_i f_i \frac{1}{y_i} - \frac{1}{G(y)}$$

Notice that the equation $\frac{1}{H(y)} = \sum_i f_i \frac{1}{y_i}$ defines H(y) the harmonic average of the $\frac{1}{y_i}$'s and that H(y) = G(y) (remember that the harmonic average is the (-1)-average while the geometric average plays the role of 0-average and that the r-average is an increasing function of r.) Therefore,

$$\frac{1}{G(y)} \frac{d}{da} g_a = \frac{1}{H(y)} - \frac{1}{G(y)} = 0$$

And thus, $\frac{d}{da} g_a = 0$ insuring that g_a is increasing and therefore $g_1 = g_0$. This also establishes that Q_p in [1] with correction factor $(1 / g_f)$ is smaller than the value without correction factor $(1 / g_0)$.

Variations

A very important point is related to optimization. What should a vendor do in order to obtain a better score? In other words, are there any guidelines to improve the perfor- mance? The question here is "how do the various averages reward decreases in the individual contributors depending on their relative size?". From the definition of the r- average

$$m_r^r = \sum_i f_i x_i^r$$

we can determine the increase in the r-average as a function of the increase in one query time assuming all the others are equal. Differentiating the above equation one obtains the following equation

$$\frac{?m_r}{?x_i} = f_i \frac{x_i^{r-1}}{m_r^{r-1}}$$

This translates into

$$dm_r = f_i \frac{x_i^{r-1}}{m_r^{r-1}} dx_i$$

which yields

$$\frac{dm_r}{m_r} / \frac{dx_i}{x_i} = \frac{f_i x_i^r}{\sum f_i x_i^r}$$

which links the relative variation of an individual query time to the variation of the measure of central tendency. The term at the left side of the above equation is referred to as the "elasticity" in Economics. Whenever $r > 0$ we have $x_1 > x_2 \_ x_1^r > x_2^r$ and therefore, the elasticity of m_r with respect to x_i is an increasing function of x_i . In other words, the bigger x_i the bigger the relative decrease of m_r for a given relative decrease $d x_i / x_i$. This is true in particular for the half-way average corresponding to r = 1/2 (and for the simple mean which we knew already). It is not true for the geometric average which treats equally large and small observations.For the displaced geometric averages the argument is similar. Starting with the definition of the a-displaced geometric average

$$\log (a + g_a) = \sum_i f_i \log (a + x_i)$$

and taking the differential in both sides assuming that x_i only varies

$$\frac{dg_a}{g_a} \frac{g_a}{a + g_a} = f_i \frac{x_i}{a + x_i} \frac{dx_i}{x_i}$$

Assuming that all the weights f_i are equal, noticing that $\frac{g_a}{a + g_a}$ is constant, and that the function x--> x/ (a + x) is monotonic increasing, it is clear that the larger x_i the larger the relative increase of g_a for a given relative increase of x_i . Therefore the a-displaced average is a balanced measure.

APPPENDIX B

As we saw, the formula for the f-displaced average is

$$\log (g_f + f) = \frac{\log (x_1 + f) + \log (x_2 + f) + ... + \log (x_n + f)}{n}$$

The quantity g_f defined above is used in the definition of the power metric when the ratio between the max and the min query time is larger than 1000; f is defined as the maximum query time divided by 1000. As shown in appendix A, g_f is always larger than g (the geometric average) no matter what the value of is as long as it is positive. This property is used to penalize a vendor who would have a "benchmark special" resulting in one query time disproportionally small compared to the other query times.

As a result, there are two formulas for the power metric in [1]. One is used when the max over min query time ratio is larger than 1000 and the other when the ratio is larger than 1000. Table 3 shows sample data illustrating the difficulties associated with this "dual formula" situation resulting in a lack of "continuity". First, notice set 2 which illustrates the point that the simple mean is very sensitive to large out of scope values and set 3 which illustrates the point that the geometric average g is very sensitive to small values. Set 4 involves a marginal case where the max/min ratio is 1000 and thus, the geometric average is used (value 2.86). In set 3, the max/min ratio is larger than 1000 and is used (value 2.88). 2.88 is very close to 2.86 but the problem is that set 3 is "better" than set 4 and yet the power (inverse of the geometric average) decreases!

TABLE 3. Variations of the Current Metric for Sample Data

	set 1	set 2	set 3	set 4	set 5
x_1	1	1000	0.001	0.01	0.01
x_2	2	2	2	2	2
x_3	3	3	3	3	3
x_4	4	4	4	4	4
x_5	5	5	5	5	5
x_6	6	6	6	6	6
x_7	7	7	7	7	7
x_8	8	8	8	8	8
x_9	9	9	9	9	9
x_{10}	10	10	10	10	10.1
mean	5.5	105.4	5.400	5.401	5.411
g	4.528	9.036	2.270	2.857	2.860
g_f			2.880		3.060

The situation is slightly different when comparing set 4 and set 5. In set 5 the max/min is higher than 1000 so g_f is used with f = 10.1/1000 = .0101 yielding the value 3.06 compared to set 4 which yields 2.86 with the geometric average. Looking at set 5 in isolation one can see that using the "corrected" metric does penalize (3.06 vs. 2.86) but comparing set 5 and set 4 there is a big drop from set 5 to set 4 (about 7%) but set 4 and set 5 are almost the same. Actually, if we had used to geometric average, we would have concluded that set 4 and set 5 where about the same (2.86). Indeed, the only real difference between set 4 and set 5 is that the formula has changed! Based on the recommendation of this author the correction formula has been abandoned by the TPC-D subcommittee.

REFERENCES

[1] TPC BENCHMARK D (Decision Support), Working Draft 6.0, August 1, 1993, Edited by François Raab, Transaction Processing Council.

[2] Calot, G., Cours de Statistique Descriptive, 1964, Dunod, Paris