Edgar H. Sibley
Panel Editor

Using the arithmetic mean to summarize normalized benchmark results leads to mistaken conclusions that can be avoided by using the preferred method: the geometric mean.

HOW NOT TO LIE WITH STATISTICS: THE CORRECT WAY TO SUMMARIZE BENCHMARK RESULTS

PHILIP J. FLEMING and JOHN J. WALLACE

In the literature, performance results are frequently summarized using the arithmetic mean of performance ratios, leading, in some cases, to wrong conclusions (see Tables 2 and 3 in [2]) or, at best, inappropriate statistics (see Tables[1] 12 and 13 in [3]). We hope to elucidate this inadvertent misuse of statistics in reporting results by pointing out why the arithmetic mean should not be used to summarize normalized performance numbers, and showing why the geometric mean is the more appropriate measure. We do this in the form of some simple rules for improved statistical analysis of performance benchmark results.

THREE RULES FOR SUMMARIZING PERFORMANCE BENCHMARK RESULTS

In comparing computers according to some metric, such as object size, run time, or throughput, it is common to run benchmarks, normalize the results to a "known machine," and then average these normalized quantities. The desire to have *one number* that represents relative system performance is certainly understandable, as we want to draw simple conclusions about one machine's worth compared to others. However, in order that these conclusions be meaningful and useful, the following three rules should be observed.

RULE 1: *Do Not Use the Arithmetic Mean to Average Normalized Numbers*

The arithmetic mean of N numbers is the sum of the numbers divided by N. When the arithmetic mean is calculated from normalized numbers, the results are meaningless. In Table I, derived from an example reported in the literature [2], all run times are normalized to Machine R. These normalized numbers are then averaged, and the conclusion drawn that Machine M is one percent slower than R, and Machine Z, seven percent slower.

However, if we normalize to Machine M and not R, (as shown in Table II), we must conclude that R is now 32 percent slower than M. *How can this be?* It cannot:

[1] The major conclusions in [3] are not invalidated by the results in this paper.

© 1986 ACM 0001-0782/86/0300-0218 75¢

The problem is the arithmetic mean, which gives meaningless results for normalized numbers.

To illustrate the problems with the arithmetic mean using a simple example, consider three machines with the benchmark run times given in Table III. Machine Y is twice as fast as Machine X for Benchmark 1, but half as fast for Benchmark 2. Similarly, Machine Z is half as fast as Machine X for 1, but twice as fast for 2. Intuitively, these three machines have equivalent performance: Each is slower than the others on one of the benchmarks, but faster by the same ratio on the other benchmark. However, if we normalize to Machine X and compute the arithmetic mean, we find that Machines Y and Z are 25 percent slower than X.

Worse yet, if we normalize to Machine Y and compute the arithmetic mean (Table IV), we find that Machine Y is now 25 percent faster than X and *more than two times faster than* Z despite the fact that the total benchmark run times for X and Z were *less than* that for Y. Clearly, the arithmetic mean is worthless in this context: When used indiscriminately, it leads to very wrong conclusions.

As a corollary to what has already been discussed, it is now relevant to introduce RULE 1.1: *The sum of normalized numbers is also meaningless.* This is fairly obvious since the sum is merely N times the arithmetic mean.

RULE 2: *Use the Geometric Mean to Average Normalized Numbers*

The geometric mean of N numbers is the product of the numbers to the $1/N$th power. Unlike the arithmetic mean, the geometric mean is meaningful when applied to normalized numbers. In Table V, the numbers from our simple XYZ example are repeated, but this time with the geometric means. Now the conclusion is more useful: The machines are shown to be roughly equal. Even if we normalize to Machine Y and not X, the results are the same (Table VI).

If, as in Table VII, we use the geometric mean instead of the arithmetic mean for the results presented in Table I, then we conclude that Machine M is 14 percent faster than Machine R, and Machine Z is 16 percent faster—very different conclusions from those presented in the original paper. The same results are derived in Table VIII, where run times are normalized to Machine M instead of Machine R, since 1.17 is the reciprocal of 0.86. Further, by comparing Machine M to Machine Z directly from the geometric means without renormalizing, we see that Machine M is 2 percent slower than Machine Z for these benchmarks ($0.86/0.84 = 1.02$). The ability to compare means without regard to normalization is an important property and unique to the geometric mean, as shown in "A Proof that the Geometric Mean Is the Only Correct Average of Normalized Measurements" section.

To summarize this discussion, we present the two corollaries to RULE 2:

RULE 2.1: *The geometric mean can be used regardless of how the numbers are normalized.*

TABLE I. Incorrect Use of the Arithmetic Mean on Normalized Numbers

Benchmark	Processor		
	R	M	Z
E	417 (1.00)	244 (0.59)	134 (0.32)
F	83 (1.00)	70 (0.84)	70 (0.85)
H	66 (1.00)	153 (2.32)	135 (2.05)
I	39,449 (1.00)	33,527 (0.85)	66,000 (1.67)
K	772 (1.00)	368 (0.48)	369 (0.45)
Arithmetic mean	(1.00)	(1.01)	(1.07)

The numbers in parentheses are normalized to Machine R.
This table has been adapted from [2] by permission of the author.

TABLE II. Same Raw Data, but Different Results

Benchmark	Processor		
	R	M	Z
E	417 (1.71)	244 (1.00)	134 (0.55)
F	83 (1.19)	70 (1.00)	70 (1.00)
H	66 (0.43)	153 (1.00)	135 (0.88)
I	39,449 (1.18)	33,527 (1.00)	66,000 (1.97)
K	772 (2.10)	368 (1.00)	369 (1.00)
Arithmetic mean	(1.32)	(1.00)	(1.08)

The numbers in parentheses are normalized to Machine M.

TABLE III. Another Incorrect Use of the Arithmetic Mean on Normalized Numbers

Benchmark	Processor		
	X	Y	Z
1	20 (1.00)	10 (0.50)	40 (2.00)
2	40 (1.00)	80 (2.00)	20 (0.50)
Arithmetic mean	(1.00)	(1.25)	(1.25)

The numbers in parentheses are normalized to Machine X.

TABLE IV. The Arithmetic Mean Is Sinking Fast

Benchmark	Processor		
	X	Y	Z
1	20 (2.00)	10 (1.00)	40 (4.00)
2	40 (0.50)	80 (1.00)	20 (0.25)
Arithmetic mean	(1.25)	(1.00)	(2.13)

The numbers in parentheses are normalized to Machine Y.

TABLE V. Correct Use of the Geometric Mean

Benchmark	Processor		
	X	Y	Z
1	20 (1.00)	10 (0.50)	40 (2.00)
2	40 (1.00)	80 (2.00)	20 (0.50)
Geometric mean	(1.00)	(1.00)	(1.00)

The numbers in parentheses are normalized to Machine X.

TABLE VI. The Geometric Mean Is Independent of Normalization

Benchmark	Processor		
	X	Y	Z
1	20 (2.00)	10 (1.00)	40 (4.00)
2	40 (0.50)	80 (1.00)	20 (0.25)
Geometric mean	(1.00)	(1.00)	(1.00)

The numbers in parentheses are normalized to Machine Y.

TABLE VII. Another Correct Use of the Geometric Mean

Benchmark	Processor		
	R	M	Z
E	417 (1.00)	244 (0.59)	134 (0.32)
F	83 (1.00)	70 (0.84)	70 (0.85)
H	66 (1.00)	153 (2.32)	135 (2.05)
I	39,449 (1.00)	33,527 (0.85)	66,000 (1.67)
K	772 (1.00)	368 (0.48)	369 (0.45)
Geometric mean	(1.00)	(0.86)	(0.84)

The numbers in parentheses are normalized to Machine R.

TABLE VIII. A Different Normalization

Benchmark	Processor		
	R	M	Z
E	417 (1.71)	244 (1.00)	134 (0.55)
F	83 (1.19)	70 (1.00)	70 (1.00)
H	66 (0.43)	153 (1.00)	135 (0.88)
I	39,449 (1.18)	33,527 (1.00)	66,000 (1.97)
K	772 (2.10)	368 (1.00)	369 (1.00)
Geometric mean	(1.17)	(1.00)	(0.99)

The numbers in parentheses are normalized to Machine M.

RULE 2.2: *The geometric mean can be used even if the numbers are not normalized; the resulting means can then be normalized.*

RULE 3: *Use the Sum (or arithmetic mean) of Raw, Unnormalized Results whenever This "Total" Has Some Meaning*

Sometimes, the sum of benchmark results has meaning: for example, total run time for a set of benchmarks. However, it is important to calculate this sum using raw, unnormalized data since we have shown that summing (or taking the arithmetic mean) of normalized numbers gives worthless results. Ratios of these unnormalized sums can then be taken to determine relative performance.

When summing or taking the arithmetic mean of raw results, it is implied that each individual benchmark is of equal importance. Typically, however, you want to weight each benchmark result to simulate a real load. In Table IX, for example, where our simple XYZ example is repeated, Benchmark 1 consumes 60 percent of our load mix, and Benchmark 2 consumes 40 percent. This means that Machine X is now 14 percent "faster" than Z and 36 percent "faster" than Y. This conclusion holds regardless of how we normalize the arithmetic mean since we have started with raw (unnormalized) data.

A PROOF THAT THE GEOMETRIC MEAN IS THE ONLY CORRECT AVERAGE OF NORMALIZED MEASUREMENTS

Earlier, we showed how using the arithmetic mean to average normalized measurements leads to inconsistencies, whereas using the geometric mean does not. In this section, we provide a proof that the geometric mean is the only average that has the multiplicative property and therefore the only appropriate measure of the mean in the present context. Although this result does not represent original mathematics (i.e., it is equivalent to Theorem 4 in [1] by a logarithmic transformation), it is presented here as a convenience to the reader.

The multiplicative property can be stated simply: The mean of the products equals the product of the means. More precisely, suppose we have N benchmarks of interest, $\beta_1, \ldots, \beta_N$, and three machines X, Y, and Z whose performance we would like to compare. After running the benchmarks on these machines, we find that β_i ran for x_i seconds on Machine X, y_i seconds on Machine Y, and z_i seconds on Machine Z. We then form the ratios $a_i = y_i/x_i$ and $b_i = z_i/y_i$. It is customary to say that Machine X runs β_i a_i times as fast as Machine Y and, similarly, that Machine Y runs β_i b_i times as fast as Machine Z. We may also conclude that Machine X runs β_i $a_i b_i$ times as fast as Machine Z and that Machine Y runs β_i a_i^{-1} as fast as Machine X. By choosing a number, say A, that summarizes the overall performance comparison between Machine X and Machine Y, we may now move to the statement that, overall, Machine X is

TABLE IX. Sums of Raw Data Can Make Sense

Benchmark	Weight	Processor		
		X	Y	Z
1	0.6	20	10	40
2	0.4	40	80	20
Weighted arithmetic mean		28	38	32
Normalized to X		1.00	1.36	1.14

A times as fast as Machine Y on $\beta_1, \ldots, \beta_N$. If we now choose a number, B, and assert that Machine Y is B times as fast as Machine Z, overall, then common sense will dictate that Machine X should be AB times as fast as Machine Z. This is the multiplicative property: The product of A and B should equal the mean of $a_1b_1, \ldots, a_Nb_N$. (For numerical examples, see Tables V, VI, and VII on page 220).

To formulate the problem mathematically, let $A = f(a_1, \ldots, a_n)$. In other words, A is some unknown function, f, of $a_1, \ldots, a_n$. We assume $a_i > 0$. Since A is an unweighted expected value or mean, the function f must satisfy the following three properties:

Property 1 (reflexive property):

$$f(a, \ldots, a) = a;$$

Property 2 (symmetric property):

$$f(a_1, \ldots, a_n) = f(a_{\sigma(1)}, \ldots, a_{\sigma(n)})$$

for all permutations σ of the numbers $1, \ldots, n$. This second property maintains that the order of the arguments of f does not affect A.

Property 3 (multiplicative property):

$$f(a_1b_1, \ldots, a_nb_n) = f(a_1, \ldots, a_n)f(b_1, \ldots, b_n).$$

We claim that Properties 1 through 3 uniquely characterize the geometric mean. To see this, first note that the geometric mean does satisfy Properties 1 through 3. We now prove that, if f satisfies Properties 1 through 3, then $f(a_1, \ldots, a_n)$ is the geometric mean.

Observe that, for any $r > 0$,

$$\begin{aligned} r &= f(r, \ldots, r) \\ &= f(r, 1, \ldots, 1)f(1, r, \ldots, 1) \cdots f(1, \ldots, 1, r) \\ &= f(r, 1, \ldots, 1)^n. \end{aligned}$$

The first equality follows from Property 1, the second is arrived at by repeated applications of Property 3, and the last is Property 2. Hence, $f(r, 1, \ldots, 1) = r^{1/n}$ for any $r > 0$. Finally, we note that Properties 2 and 3, together with the above calculation, imply that

$$\begin{aligned} &f(a_1, \ldots, a_n) \\ &= f(a_1, 1, \ldots, 1)f(1, a_2, 1, \ldots, 1) \cdots f(1, \ldots, 1, a_n) \\ &= \prod_{i=1}^{n} f(a_i, 1, \ldots, 1) \\ &= \left(\prod_{i=1}^{n} a_i\right)^{1/n}. \end{aligned}$$

It can now be seen that the only choice of A that satisfies Properties 1 through 3 is the geometric mean. As a final remark, note that a weighted geometric mean, which also satisfies the multiplicative property, may be calculated as follows: Let $w_1, \ldots, w_N$ be weights such that $w_1 + \cdots + w_N = 1$. The weighted mean is then

$$\prod_{i=1}^{N} a_i^{w_i}.$$

The unweighted mean is the case $w_i = 1/N$, $i = 1, \ldots, N$.

CONCLUSIONS

In this article, we have demonstrated why the geometric mean is appropriate for summarizing normalized benchmark results, and why the arithmetic mean, when used in this context, leads to grossly incorrect conclusions.

However, it should be made clear that any measure of the mean value of data is misleading when there is large variance. For this reason, we feel that any meaningful summary of data should include some mention of the minimum and maximum of the data as well as the mean. This provides guaranteed upper and lower bounds on the relative performance with respect to the chosen set of benchmarks.

REFERENCES

1. Aczel, J. *Functional Equations*. Academic Press, New York, 1966, p. 239. A comprehensive textbook on functional equations.
2. Heath, J.L. Re-evaluation of RISC I. *Comput. Archit. News 12*, 1 (Mar. 1984), 3–10. Performance comparison of RISC versus CISC.
3. Patterson, D.A., and Sequin, C.H. A VLSI RISC. *Computer 15*, 9 (Sept. 1982), 8–21. The landmark paper formally introducing the RISC approach to computer architecture.

CR Categories and Subject Descriptors: C.4 [**Performance of Systems**]: *measurement techniques, performance attributes*
General Terms: Measurement, Performance
Additional Key Words and Phrases: benchmarking, geometric mean

Received 5/85; accepted 8/85

Authors' Present Addresses: Philip J. Fleming, AT&T Information Systems, 1100 East Warrenville Road, Naperville, IL 60566; John J. Wallace, The Foxboro Company, Foxboro, MA 02035; Electronic mail: foxvax5!jjw.