Guesstimation: A New Justification of the Geometric Mean Heuristic

Olga Kosheleva[1] and Vladik Kreinovich[2]
[1]Department of Teacher Education
[2]Department of Computer Science
University of Texas at El Paso
500 W. University
El Paso, TX 79968, USA
olgak@utep.edu, vladik@utep.edu

Abstract

In many practical situations in which the only information we have about the quantity x is that its value is within an interval $[\underline{x}, \overline{x}]$, a reasonable estimate for this quantity is the geometric mean of the bounds $\sqrt{\underline{x} \cdot \overline{x}}$. In this paper, we provide a new justification for this geometric mean heuristic.

Formulation of the problem. In many practical applications, we only know the bounds on a quantity x, and we need to make a reasonable estimate based on these bounds. In other words, if we only know the interval $[\underline{x}, \overline{x}]$ of possible values of x, and we want to select a single reasonable estimate for x, which value from the interval $[\underline{x}, \overline{x}]$ shall we select?

Due to the practicality of this problem, it is important to be able to select a reasonable estimate, and it is also important to teach students how to make these estimations.

It turns out that in many practical situations, the geometric mean $\sqrt{\underline{x} \cdot \overline{x}}$ is a reasonable estimate for x; see, e.g., [1, 5]. How can we explain the success of this heuristic?

What we do in this paper. In this paper, we provide a new justification for this geometric mean heuristic.

Example. Before we proceed with a new justification, let us give one example where this heuristic works well. This example is related to measurements; see, e.g., [2].

Measurements are never 100% accurate. In general, the result $\widetilde{x}$ of measuring the value of a physical quantity x is different from the actual (unknown) value

of this quantity. In other words, the measurement error $\Delta x \stackrel{\text{def}}{=} \widetilde{x} - x$ is, in general, different from 0.

Usually, there are many different sources of the measurement error. Let us denote by n the number of such sources. Then, the overall measurement error Δx can be described as the sum of the measurement errors coming from these n sources:

$$\Delta x = \delta x_1 + \ldots + \delta x_n.$$

Often, we do not have much information about the measurement error components δx_i; we only know the upper bound δ on the values of these components, i.e., we only know that $|\delta x_i| \le \delta$ for all i.

What is the resulting estimate Δ for the absolute value $|\Delta x|$ of the overall measurement error Δx? The worst-case estimate $\overline{\Delta}$ comes from the case when each of the components δx_i attains its largest possible value δ. In this case, the overall measurement error attains the largest possible value $\overline{\Delta} = n \cdot \delta$.

What is the smallest possible value $\underline{\Delta}$ of the (absolute value of the) overall measurement error? From the purely mathematical viewpoint, we can have $\delta x_i = 0$ for all i and thus, $\Delta x = 0$. However, as we have mentioned, from the physical viewpoint, the situation with no measurement errors is not realistic at all. In other words, at least of the error components has to be non-zero, and if there are many of these components, at least one of them should take the value close to its upper bound δ. Thus, a natural lower bound comes from the situation when one of the error components has the value $\delta x_i = \delta$ and all the others are 0s. In this case, we have $\Delta x = \delta$, so we set $\underline{\Delta} = \delta$.

Both the upper bound $\overline{\Delta}$ and the lower bound $\underline{\Delta}$ are not very probable:

- the lower bound $\underline{\Delta}$ requires that many error components are almost zeros, while
- the upper bound $\overline{\Delta}$ requires that all the error components are as large as possible.

A reasonable estimate should be somewhere in between.

How can we get such a reasonable estimate? Since we do not have any information about the probabilities of different measurement errors $\delta_i \in [-\Delta, \Delta]$ or about the dependence between different measurement errors, it is reasonable to assume that every error component is uniformly distributed on the corresponding interval, and that different error components are independent random variables. Under this assumption, the overall measurement error is the sum of n identically distributed independent random variables δx_i. It is known that, due to the Central Limit Theorem (see, e.g., [3]), for large n, the distribution of this sum is close to Gaussian.

To describe this Gaussian distribution, we need to know its sum and its variance. The mean of the sum of several independent random variables is the sum of their means, and the variance is equal to the sum of their variances. For each variable δ_i uniformly distributed on the interval $[-\delta, \delta]$, the mean is 0, and

the variance is equal to $\frac{1}{3} \cdot \delta^2$. Thus, the sum Δx is normally distributed with 0 mean and variance $V = n \cdot \frac{1}{3} \cdot \delta^2$.

With a high probability of 99.9%, we can conclude that the actual value Δx differs from the mean 0 by no more than three standard deviations $3 \cdot \sigma = 3 \cdot \sqrt{V}$. Thus, the reasonable upper bound Δ on $|\Delta x|$ is the value

$$\Delta = 3 \cdot \sqrt{n \cdot \frac{1}{3} \cdot \delta^2} = \sqrt{3} \cdot \sqrt{n} \cdot \delta.$$

Almost exactly this same value comes from applying the geometric mean heuristic to the values $\underline{\Delta} = \delta$ and $\overline{\Delta} = n \cdot \delta$:

$$\Delta \approx \sqrt{\underline{\Delta} \cdot \overline{\Delta}} = \sqrt{\delta \cdot (n \cdot \delta)} = \sqrt{n} \cdot \delta.$$

Formalization of the general problem. We want to be able, given the bounds $\underline{x}$ and $\overline{x}$, to produce a value $x \in [\underline{x}, \overline{x}]$ which depends on these bounds. Let us depend an estimate corresponding to these bounds by $f(\underline{x}, \overline{x})$. In terms of this notation, we must find a function $f(\underline{x}, \overline{x})$ which has the property

$$\underline{x} \le f(\underline{x}, \overline{x}) \le \overline{x} \tag{1}$$

for all $\underline{x}$ and $\overline{x}$. Thus, we arrive at the following definition.

Definition 1. *By an* estimation function *we mean a function $f(\underline{x}, \overline{x})$ which is defined for all pairs of real numbers $\underline{x} > 0$ and $\overline{x} > 0$ for which $\underline{x} \le \overline{x}$ and which satisfies the property for all $\underline{x} \le \overline{x}$:*

$$\underline{x} \le f(\underline{x}, \overline{x}) \le \overline{x}. \tag{2}$$

First requirement: scale-invariance. Our first requirement is related to the fact that the numerical values of a physical quantity usually depends on the choice of the measuring unit. For example, the same distance between the two cities has different numerical values when expressed in miles or in kilometers.

Both values represent the exact same distance – or, in the general case, the same quantity. It is therefore reasonable to require that our estimate does not depend on the choice of the measuring unit. In other words, our estimate $f(\underline{x}, \overline{x})$ should be *scale-invariant* in the sense that it should not depend on the scale used for measuring the corresponding quantity.

Let us describe this requirement in precise terms. Let us assume that $\underline{x}$ and $\overline{x}$ described the numerical values of the lower and upper bounds for x as expressed in the original units. In this case, the estimate for x takes the form $f(\underline{x}, \overline{x})$.

Instead of using the original measuring unit, for every real number $\lambda > 0$, we can select a new unit which is λ times smaller than the original one. In the new

units, all numerical values are λ times larger than in the original units. Thus, in the new units, the lower and upper bounds on the quantity x takes the form $\lambda \cdot \underline{x}$ and $\lambda \cdot \overline{x}$. Thus, when applied to the bounds as described in the new units, our estimate f leads to the value $f(\lambda \cdot \underline{x}, \lambda \cdot \overline{x})$.

Our requirement is that this estimate should be the same as the original estimate $f(\underline{x}, \overline{x})$, but expressed in the new units. Describing the estimate $f(\underline{x}, \overline{x})$ in the new units means multiplying its numerical value by λ, which results in the new numerical value $\lambda \cdot f(\underline{x}, \overline{x})$. Our requirements means that this value must coincide with the estimate $f(\lambda \cdot \underline{x}, \lambda \cdot \overline{x})$ obtained by applying our heuristic to the new units.

Thus, we arrive at the following definition.

Definition 2. *We say that an estimation function $f(\underline{x}, \overline{x})$ is* scale-invariant *if for every $\underline{x} \le \overline{x}$ and for every $\lambda > 0$, we have*

$$f(\lambda \cdot \underline{x}, \lambda \cdot \overline{x}) = \lambda \cdot f(\underline{x}, \overline{x}). \tag{3}$$

Second requirement: dependence symmetry. The second requirement is related to the fact that many physical properties are not, strictly speaking, directly measurable, they describe the relation between two different quantities. Let us give a few examples to explain what we have in mind:

- The mass m describes how the force f depends on the acceleration a: $f = m \cdot a$.
- The density ρ describes how the mass m depends on the volume V:

$$m = \rho \cdot V.$$

- The velocity v describes how the distance d depends on time: $d = v \cdot t$.
- The electric resistance R describes how the voltage V depends on the current I: $V = I \cdot R$.

More generally, in many cases, the quantity x describes how a quantity x_1 depends on a quantity x_2: $x_1 = x \cdot x_2$. Once we know the value of x_2, this formula enables us to find the corresponding value of x_1.

In some case, we face an opposite problem: we know x_1 and we want to know x_2. In this case, the same physical law can be described as a dependence of x_2 on x_1: $x_2 = y \cdot x_1$, where $y \stackrel{\text{def}}{=} \dfrac{1}{x}$.

For example, if we know the travel time t, then, to find the distance d, we can use the formula $d = v \cdot t$. Vice versa, if we know the distance, then, to find the time, we can use the formula $t = s \cdot d$, where the value $s \stackrel{\text{def}}{=} \dfrac{1}{v}$ is called *slowness*; see, e.g., [4].

Similarly, if we know the current I, then we can determine the voltage V by using the formula $V = R \cdot I$. Vice versa, if we know the voltage, then, to find the current, we can use the formula $I = G \cdot V$, where $G \stackrel{\text{def}}{=} \dfrac{1}{R}$ is called *conductance*.

In general, it is reasonable to require that the estimation procedure $f(\underline{x}, \overline{x})$ is *symmetric* in the sense that it leads to the exact same result whether we apply it to the original quantity x or to the reciprocal quantity $y = \dfrac{1}{x}$. Let us formulate this requirement in precise terms.

In the original form, the bounds on x are $\underline{x}$ and $\overline{x}$, so the procedure f results in the value $x = f(\underline{x}, \overline{x})$.

The new quantity $y = \dfrac{1}{x}$ is a decreasing function of x. Thus, the smallest possible value $\underline{y}$ of y is attained when x takes the largest possible value $x = \overline{x}$, and the largest possible value $\overline{y}$ of y is attained when x takes the smallest possible value $x = \underline{x}$. In other words, $\underline{y} = \dfrac{1}{\overline{x}}$ and $\overline{y} = \dfrac{1}{\underline{x}}$. For these new bound, the procedure f leads to the estimate

$$y = f(\underline{y}, \overline{y}) = f\left(\frac{1}{\overline{x}}, \frac{1}{\underline{x}}\right).$$

This estimate corresponds to the value $x = \dfrac{1}{y}$. Our dependence symmetry requirement means that this value x must coincide with the original estimate. Thus, we arrive at the following definition.

Definition 3. *We say that an estimation function $f(\underline{x}, \overline{x})$ is* dependence-symmetric *if for every $\underline{x} \le \overline{x}$, we have*

$$\frac{1}{f\left(\dfrac{1}{\overline{x}}, \dfrac{1}{\underline{x}}\right)} = f(\underline{x}, \overline{x}). \tag{4}$$

Proposition. *There exists one and only one estimation function which is both scale-invariant and dependence-symmetric: the function $f(\underline{x}, \overline{x}) = \sqrt{\underline{x} \cdot \overline{x}}$.*

Discussion. Thus, we have indeed justified the geometric mean heuristic.

Proof. One can easily check that the geometric mean function $f(\underline{x}, \overline{x})$ is both scale-invariant and dependence-symmetric. Thus, to complete the proof, it is sufficient to prove that if an estimation function $f(\underline{x}, \overline{x})$ is scale-invariant and dependence-symmetric, then it coincides with the geometric mean.

Indeed, since the function $f(\underline{x}, \overline{x})$ is scale-invariant, it satisfies the property (3) for all $\underline{x}$, $\overline{x}$, and λ. Dividing both sides of this equality by λ, we conclude

that

$$f(\underline{x},\overline{x}) = \frac{1}{\lambda}\cdot f(\lambda\cdot\underline{x},\lambda\cdot\overline{x}). \tag{5}$$

This equality holds for every λ. In particular, for $\lambda = \dfrac{1}{\underline{x}}$, we conclude that

$$f(\underline{x},\overline{x}) = \underline{x}\cdot f\left(1,\frac{\overline{x}}{\underline{x}}\right). \tag{6}$$

We now want to use the dependence symmetry property. For the property $y = 1/x$, we similarly have

$$f(\underline{y},\overline{y}) = \underline{y}\cdot f\left(1,\frac{\overline{y}}{\underline{y}}\right). \tag{7}$$

Substituting $\underline{y} = \dfrac{1}{\overline{x}}$ and $\overline{y} = \dfrac{1}{\underline{x}}$ into this formula, we conclude that

$$f\left(\frac{1}{\overline{x}},\frac{1}{\underline{x}}\right) = \frac{1}{\overline{x}}\cdot f\left(1,\frac{1/\underline{x}}{1/\overline{x}}\right), \tag{8}$$

or, equivalently,

$$f\left(\frac{1}{\overline{x}},\frac{1}{\underline{x}}\right) = \frac{1}{\overline{x}}\cdot f\left(1,\frac{\overline{x}}{\underline{x}}\right). \tag{9}$$

Thus,

$$\frac{1}{f\left(\dfrac{1}{\overline{x}},\dfrac{1}{\underline{x}}\right)} = \frac{\overline{x}}{f\left(1,\dfrac{\overline{x}}{\underline{x}}\right)}. \tag{10}$$

Because of (6) and (10), the dependence symmetry condition (4) takes the form

$$\underline{x}\cdot f\left(1,\frac{\overline{x}}{\underline{x}}\right) = \frac{\overline{x}}{f\left(1,\dfrac{\overline{x}}{\underline{x}}\right)}. \tag{11}$$

Moving terms $f\left(1,\dfrac{\overline{x}}{\underline{x}}\right)$ into the left-hand side and all other terms into the right-hand side, we conclude that

$$f^2\left(1,\frac{\overline{x}}{\underline{x}}\right) = \frac{\overline{x}}{\underline{x}} \tag{12}$$

and therefore, that

$$f\left(1,\frac{\overline{x}}{\underline{x}}\right) = \sqrt{\frac{\overline{x}}{\underline{x}}}. \tag{13}$$

Substituting this expression (13) into the formula (6), we get

$$f(\underline{x},\overline{x}) = \underline{x}\cdot\sqrt{\frac{\overline{x}}{\underline{x}}} = \sqrt{\underline{x}\cdot\overline{x}}. \tag{14}$$

The proposition is proven.

Acknowledgments. This work was supported in part by NSF grant HRD-0734825 and by Grant 1 T36 GM078000-01 from the National Institutes of Health.

References

[1] H. S. Left, A review of [5], *Physics Today*, 2009, Vol. 62, No. 2 (February), p. 62.

[2] S. Rabinovich, *Measurement Errors and Uncertainties: Theory and Practice*, American Institute of Physics, New York, NY, 2005.

[3] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman & Hall/CRC Press, Boca Raton, Florida, 2007.

[4] S. Stein and M. Wysession, *An Introduction to Seismology, Earthquakes and Earth Structure*, Blackwell Publishing, Malden, Massachusetts, 2003.

[5] L. Weinstein and J. A. Adam, *Guesstimation: Solving the World's Problems on the Back of a Cocktail Napkin*, Princeton University Press, Princeton, New Jersey, 2008.