Why Dependence of Productivity on Group Size Is Log-Normal

Francisco Zapata, Olga Kosheleva and Vladik Kreinovich
University of Texas at El Paso
500 W. University
El Paso, TX 79968, USA
fazg74@gmail.com, olgak@utep.edu, vladik@utep.edu

Abstract

Empirical analysis shows that, on average, the productivity of a group log-normally depends on its size. The current explanations for this empirical fact are based on reasonably complex assumptions about the human behavior. In this paper, we show that the same conclusion can be made in effect, from first principles, without making these complex assumptions.

1 Formulation of the Problem

How productivity changes with group size: a qualitative description. Productivity of a group depends on its size. Usually, if first a single person has been working on a project, and then a second person arrives to help, the project speeds up. Addition of the third helper also increases the productivity, etc.

However, this is only true until a certain threshold is reached. After that, too many people just confuse each other – whether it is too many people trying to help wash the dishes after a meal, or too many programmers trying to work on a joint software project.

How productivity changes with group size: a quantitative description. Several studies analyzed how exactly productivity c changes with the group size n; see, e.g., [1, 4], and concluded that this dependence can be well described by the log-normal formula

$$c(n) = \text{const} \cdot \exp\left(-\frac{(\ln(n) - \mu)^2}{2\sigma^2}\right) \tag{1}$$

for appropriate values μ and σ.

Why this is interesting. If we know how the productivity depends on the group size, we will be able to select the group size that leads to the optimal productivity.

If we know the general formula with unknown parameters – like the above log-normal expression – then we can use a few observations corresponding to different group sizes n to find the corresponding parameters and thus, to find the optimal group size.

Existing explanations. How can we explain this interesting empirical fact? In [1, 4], this fact is explained by invoking a rather complex description of "information foraging". Under proper additional assumptions, this description indeed leads to the log-normal dependence.

Remaining problem. While the existing explanation is reasonable, it uses a lot of complex assumptions. A natural question is: are these complex assumptions really necessary – or we can derive the log-normal dependence without them, from first principles?

What we do in this paper. In this paper, we show that the empirical log-normal dependence can indeed be derived from first principles, without the need to invoke additional complex assumptions.

2 Our Explanation

Towards an explanation: main idea. Our main idea is that, unless a group consists of two people, there is always some structure in this group: some pairs of people collaborate (somewhat) more, other pairs collaborate (somewhat) less. Crudely speaking, this means that within the group there is a kind of a hierarchy:

- the group as a whole can be divided into subgroups – so that within each subgroup, there is a higher degree of collaboration, while between people from different subgroups, the level of collaboration is smaller;
- each subgroup, in its turn, can be subdivided into sub-sub-groups so that within each such sub-sub-group, people collaborate slightly more, etc.

How to transform this idea into a precise description. Let us start with the units (sub-...-groups) of the smallest size. Let $c_1(n_1)$ denote the average productivity of such a unit when its size is n_1.

When several such units start working together, their productivity increases. If we have n_2 units working together, then the original productivity $c_1(n_1)$ is increased by some factor depending on n_2. Let us denote this factor by $c_2(n_2)$. In this case, the resulting overall productivity is equal to the product $c_1(n_1) \cdot c_2(n_2)$. Together, n_2 units with, on average, n_1 persons in each contain $n_1 \cdot n_2$ people.

When the number of people is large enough, these are several such "second-order" units of size $n_1 \cdot n_2$. Let us denote the average number of such second-order units by n_3. When we have n_3 units working together, then the original productivity $c_1(n_1) \cdot c_2(n_2)$ of each second-order unit is increased by some factor depending on n_3. Let us denote this factor by $c_3(n_3)$. In this case, the resulting

overall productivity is equal to the product $c_1(n_1) \cdot c_2(n_2) \cdot c_3(n_3)$. Together, n_3 units with, on average, $n_1 \cdot n_2$ persons in each contain $n_1 \cdot n_2 \cdot n_3$ people.

We can repeat this construction again and again. If we already considered k-th order units with productivity $c_1(n_1) \cdot c_2(n_2) \cdot \ldots \cdot c_k(n_k)$, then we can also consider a situation in which we have several such units. Let us denote the average number of such k-order units working together by n_{k+1}. When we have n_{k+1} units working together, then the original productivity

$$c_1(n_1) \cdot c_2(n_2) \cdot \ldots \cdot c_k(n_k)$$

of each k-order unit is increased by some factor depending on n_{k+1}. Let us denote this factor by $c_{k+1}(n_{k+1})$. In this case, the resulting overall productivity is equal to the product

$$c_1(n_1) \cdot c_2(n_2) \cdot \ldots \cdot c_k(n_k) \cdot c_{k+1}(n_{k+1}).$$

Together, n_{k+1} units with, on average,

$$n_1 \cdot n_2 \cdot \ldots \cdot n_k$$

persons in each contain

$$n_1 \cdot n_2 \cdot \ldots \cdot n_k \cdot n_{k+1}$$

people.

In general, if we have ℓ hierarchical levels, then for the resulting group of

$$n = n_1 \cdot \ldots \cdot n_\ell, \tag{2}$$

the corresponding productivity is equal to the product

$$c(n) = c_1(n_1) \cdot \ldots \cdot c_\ell(n_\ell). \tag{3}$$

Let us simplify these formulas. Formulas containing several consequent multiplications can usually be simplified if we consider logarithms instead of the original values. Logarithm of a product is equal to the sum of the logarithms, and sums are easier to compute and easier to analyze than products – it is worth mentioning that this simplification is exactly what logarithms were originally invented for.

In particular, from the formula (2), we get

$$N = N_1 + \ldots + N_\ell, \tag{4}$$

where we denoted $N \stackrel{\text{def}}{=} \ln(n)$ and $N_i \stackrel{\text{def}}{=} \ln(n_i)$.

In terms of these new variables, we have $n = \exp(N)$ $n_i = \exp(N_i)$, and so, $c(n) = C(N)$ and $c_i(n_i) = C_i(N_i)$, where we denoted $C(N) \stackrel{\text{def}}{=} c(\exp(n))$ and $C_i(N_i) \stackrel{\text{def}}{=} c_i(\exp(n_i))$. So, we conclude that

$$C(N) = C_1(N_1) + \ldots + C_\ell(N_\ell), \text{ where } N = N_1 + \ldots + N_\ell. \tag{5}$$

What can we deduce from this simplified formula? We are interested in finding the productivity of a group of given size n, and thus, of given value of the logarithm $N = \ln(n)$. This value N is all we know, we do not know a priori what the values $N_1, \ldots, N_\ell$ are.

Since we do not know the values N_i, it is reasonable to consider them as random variables – and take an average over all possible combinations of these values. We have no reason to believe that some combinations of values are more probable than others. So, following Laplace's Indeterminacy Principle, it is reasonable to assume that all these combinations have the same probability; see, e.g., [3]. So, the resulting estimate for $C(N)$ is the average of all the sums

$$C_1(N_1) + \ldots + C_\ell(N_\ell)$$

over all combinations $N_1, \ldots, N_\ell$ for which

$$N_1 + \ldots + N_\ell = N.$$

The average is proportional to the sum, so we have

$$C(N) \sim \sum_{N_1+\ldots+N_\ell} C_1(N_1) + \ldots + C_\ell(N_\ell). \tag{6}$$

Let us simplify some more. Usually, we can simplify computations if we replace the discrete sum by its continuous approximation – the corresponding integral. This is how statistical physics works, when instead of considering a difficult-to-analyze collection of 10^{23} atoms, we consider an easier-to-analyze continuous medium; see, e.g., [2]. This is how Stirling's formula for approximating the factorial $n! = 1 \cdot 2 \cdot \cdot n$ can be derived: we take a logarithm of both sides, and then approximate the resulting sum $\ln(n!) = \ln(1) + \ldots + \ln(n)$ with the corresponding integral

$$\int_1^n \ln(x)\, dx = (x \cdot \ln(x) - x)|_1^n = n \cdot \ln(n) - n + 1,$$

hence

$$n! = \exp(\ln(n!)) \approx \exp(n \cdot \ln(n) - n + 1) =$$
$$\exp(n \cdot \ln(n)) \cdot \exp(-n) \cdot e = \left(\frac{n}{e}\right)^n \cdot \text{const}.$$

For the above sum, the corresponding integral takes the form

$$C(N) \sim \int C_1(N_1) \cdot \ldots \cdot C_\ell(N_\ell)\, dN_1 \ldots\, dN_\ell, \tag{7}$$

where the integral is taken over all the tuples $(N_1, \ldots, N_\ell)$ for which

$$N_1 + \ldots + N_\ell = N.$$

This implies log-normality. The above integral is nothing else but a convolution of ℓ functions $C_1(N_1)$, ..., $C_\ell(N_\ell)$. This is easy to see when $\ell = 2$: in this case, $N_2 = N - N_1$, and the integral takes the usual convolution form

$$\int C_1(N_1) \cdot C_2(N - N_1)\, dN_1.$$

Similarly, we can see that it is convolution for all ℓ.

In statistics, it is known that under some reasonable conditions, the approximation of the sum of a large number of reasonably small independent random variables is close to Gaussian. This statement is known as the *Central Limit Theorem*; see, e.g., [5].

When we have two independent random variables, with probability density functions $\rho_1(x_1)$ and $\rho_2(x_2)$, then the probability density function $\rho(x)$ of their sum $x = x_1 + x_2$ is the convolution of the corresponding probability density functions $\rho(x) = \int \rho(x_1) \cdot \rho_2(x - x_1)\, dx_1$. Similarly, when we have ℓ independent random variables, with probability density functions $\rho_1(x_1)$, ..., $\rho_\ell(x_\ell)$, then the probability density function $\rho(x)$ of their sum $x = x_1 + \ldots + x_\ell$ is the convolution of the corresponding probability density functions

$$\rho(x) = \int \rho(x_1) \cdot \ldots \cdot \rho_\ell(x_\ell)\, dx_1 \ldots\, dx_\ell,$$

where the integration is over all the tuples $(x_1, \ldots, x_\ell)$ for which $x_1 + \ldots + x_\ell = x$.

So, in terms of probability density functions, the Central Limit Theorem states that, under reasonable conditions, if we have a large number of non-negative functions, then their convolution is close to the probability density function of the Gaussian distribution

$$\rho(x) = \text{const} \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

for appropriate values μ and σ. (Of course, we need to *normalize* the corresponding non-negative functions, to make sure that their integral is equal to 1 and thus, that they can be interpreted as probability density functions.)

Thus, due to the formula (7), the dependence $C(N)$ of productivity c on the logarithm $N = \ln(n)$ of the group size has the Gaussian form. And this is exactly the definition of the log-normal dependence on n: when there is a Gaussian dependence on $\ln(n)$.

So, we indeed derived log-normality of the dependence $c(n)$ from first principles.

Acknowledgments

This work was supported in part by the National Science Foundation grants HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and

DUE-0926721, and by an award "UTEP and Prudential Actuarial Science Academy and Pipeline Initiative" from Prudential Foundation.

The authors are thankful to Tanmay Bhowmik for valuable discussions.

References

[1] T. Bhowmik, N. Niu, W. Wang, J.-R. C. Cheng, and X. Caop, "Optimal group size for software change tasks: a special information foraging perspective", *IEEE Transactions on Cybernetics*, to appear.

[2] R. Feynman, R. Leighton, and M. Sands, *The Feynman Lectures on Physics*, Addison Wesley, Boston, Massachusetts, 2005.

[3] E. T. Jaynes and G. L. Bretthorst, *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, UK, 2003.

[4] P. Pirolli, "An elementary social information foraging model", *Proceedings of the 27th 2009 Conference on Human Factors in Computing Systems CHI'2009*, Boston, Masssachusetts, April 4–9, 2009, pp. 605–614.

[5] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman & Hall/CRC, Boca Raton, Florida, 2011.