



Natural distribution

Tiia Grönholm, Arto Annila *

Department of Physical Sciences, Institute of Biotechnology and Department of Biological and Environmental Sciences, FI-00014, University of Helsinki, Finland

Received 1 March 2007; received in revised form 11 July 2007; accepted 24 July 2007

Available online 6 August 2007

Abstract

Log-normal distributions describe data from diverse disciplines of science. However, the fundamental basis of log-normal distributions is unknown. We suggest that the skewed distributions are outcomes of natural processes i.e. they result from the principle of increasing entropy. Fluctuations during the course of evolution toward more probable states yield multiplicative variations about the mean. The non-linear dispersion of thermodynamic states, i.e. matter and energy defined by chemical potentials, underlies the skew. Cumulative curves of skewed distributions without integrable analytical forms are characteristic of natural processes.

© 2007 Elsevier Inc. All rights reserved.

Keywords: Entropy; Evolution; Log-normal distribution; Natural process; Skewed distribution

1. Introduction

Skewed distributions [1,2] describe variable variations in many fields of science [3–5] ranging, e.g. from ecology to economy and from medicine to material sciences. For example, variations in animal and plant species just as in incomes appear log-normal, i.e. normal when presented as a function of logarithm of the variable. Dose–response relations just as grain sizes from grinding processes show log-normal distributions. Also a variable, such as atmospheric aerosol size,

* Corresponding author. Tel.: +358 9 191 50629; fax: +358 9 191 50639.
E-mail address: arto.annila@helsinki.fi (A. Annila).

may display more than one log-normal modes that relate to distinct processes [6,7]. The cumulative curves of skewed distributions are characteristic of organism growth and population dynamics as well as epidemic courses.

The probability density function [8,9] is defined for the logarithm of a normal random variable x

$$f(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp \left[-\frac{(\ln x - \mu)^2}{2\sigma^2} \right], \quad (1)$$

where μ and σ are the median and standard deviation. Typically a variable that is a result of many small independent factors as a multiplicative product can be modeled as log-normal. The characteristic skew of log-normal distributions is understood to arise from fluctuations in successive steps that are subject to non-linear confines [10]. The law of proportionate effect, i.e. the change in the variable at any step of the process is a random proportion of the previous value, underlies log-normal distributions. Especially biological processes generate log-normal data [11,12]. Skewed distributions are common when variance is large. Indeed ‘exceptional’ values may indicate a multiplicative normal distribution rather than anomalies, often rejected from an analysis predestined to comply with an additive normal distribution.

The log-normal distribution, in spite of its mathematical simplicity and universality, has not been derived from fundamental grounds. Of course the log-normal distribution does not have to be the exact description but an effective approximation for diverse variable variations found in nature [12]. In this study, we seek for the general basis for skewed distributions from *natural processes* that refer to the particular processes where entropy increases. Our aim is motivated by the words of Jacobus C. Kapteyn, who laid the foundation of the theory of frequency curves [2] as early as 1903 – the connection between the form of the curve and the causes that yield it, may lead us to precious indications about the nature of the true causes.

2. Natural processes

The exponential transformation from a log-normal to a normal distribution directs us to look for the basis of log-normal distribution from exponential stationary-state partitions that correspond to the maximum entropy [13]. Natural processes, e.g. molecular diffusion, heat conduction and chemical reactions, evolve towards more probable states, i.e. increase in entropy by decreasing free energy using various mechanisms of energy transduction [14,15]. Since $x \geq 1$ we reason that the log-normal distribution stem from physical processes with conserved positive quantities, e.g. matter.

We consider log-normal distributions to arise from natural processes on the basis of recently formulated statistics of open systems [15] that underlies their thermodynamics. The statistics relates a state of system, i.e. distribution of constituents, with a probability P . The objective is, according to the principle of increasing entropy $dS > 0$, to find the distribution corresponding to the maximum probability. The entropy $S = R \ln P$ of an open system is [15]

$$\ln P \approx \frac{1}{RT} \sum_{j=1} N_j \left(\sum_k \mu_k + \Delta Q_{jk} - \mu_j + RT \right), \quad (2)$$

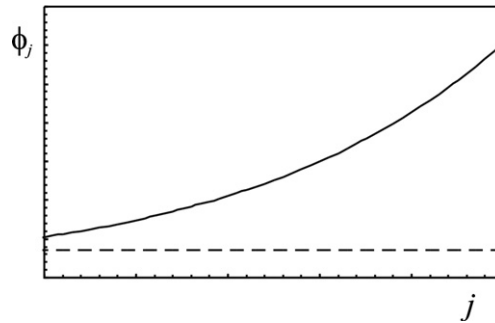


Fig. 1. Non-linear densities-in-energy ϕ_j identify to the energy contents of entities. An entity in class j is viewed to have assembled from base units of matter N_1 and quanta of energy ε_1 . The associated density-in-energy $\phi_j = N_1^j \exp(j\varepsilon_1)$ converts to a linear scale (dashed) by a transformation $x_j = \ln\phi_j/j$ that is used to display a skewed distribution as a normal distribution on a logarithmic scale.

where variables N_j enumerate various entities of the system in classes j . When entities interact with each other the system evolves toward more probable states by energy transduction that diminishes the free energy. For example, molecules are entities of systems known as cells, and cells are entities of systems known as organisms, and organisms are entities of ecosystems. Matter and energy in chemical and biological aggregates e.g. molecules, cells, organism, etc., are characterized by exponential densities-in-energy [16] $\phi_j = N_j \exp(G_j/RT)$, usually given by logarithms i.e. chemical potentials [17] $\mu_j = RT \ln[N_j \exp(G_j/RT)]$ where G_j is the Gibbs free energy relative to the thermal energy RT e.g. per mole (Fig. 1). External energy ΔQ_{jk} may also couple to aggregate assembly processes, such as to a growth of an organism or a population. Disintegration processes that dissipate energy from the system may happen spontaneously or induced using external energy. Thermodynamics obviously limits also constructions and destructions by man. The Eq. (1) expresses these thermodynamics imperatives that all aggregate systems have to comply with.

The maximum probability distribution has been attained when $d(\ln P) = 0$. At the stationary state all potentials across all jk -transitions (reactions) are equal [15]

$$\begin{aligned}
 d(\ln P) = 0 &\iff \mu_j = \sum_k \mu_k + \Delta Q_{jk} \\
 &\iff N_j = \prod_k N_k \exp [(\Delta Q_{jk} - \Delta G_{jk})/RT].
 \end{aligned}
 \tag{3}$$

The most probable distribution of entities is exponential in agreement with the law of mass action [18]. Ingredients N_k in the form of a product shows that the outcome N_j is a result of k -multiplicative operations. The stationary non-equilibrium state is maintained by the external energy, in other words determined by external conditions in agreement with LeChatelier’s principle.

Since every entity is composed of base units N_1 , e.g. atoms, densities-in-energy can be expressed recursively

$$\begin{aligned}
 \phi_j &= N_j \exp(\varepsilon_j) = N_{j-1} N_1 \exp(\varepsilon_{j-1} + \varepsilon_1) \\
 &= N_1^j \exp(j\varepsilon) = \exp(j \ln \phi_1)
 \end{aligned}
 \tag{4}$$

where we have used a shorthand notation $\varepsilon_{jk} = (\Delta Q_{jk} - \Delta G_{jk})/RT$ for energy contributions that govern occupancies of each class. The last expression is simplified in the statistical sense so that the addition of each base unit of matter to the aggregate entity, e.g. a growing individual, would involve a quantum of energy ε . Obviously strengths of chemical bonds depend on the particular atoms involved however the approximation is justified since, e.g. all biological organisms are on the average assembled from the same basic constituents.

3. Stationary state distributions

A distribution serves to distinguish entities from each other by a property that relates to a parameter value. In general the choice of property under an examination is inherently a subjective decision. In other words a distribution not only displays the parameter variation but also reflects the criteria to accept an entity to the distribution. A common choice is to include in a distribution entities from adjacent classes. These differ according to Eq. (3) by the number of k -multiplicative operations. Certainly, not only the number of steps but also the amount of matter and energy involved in each step of assembly could vary. However for quantized matter and energy, a step with variation is in fact a varying sequence of smaller steps. Hence the adjacent classes differ only by one unit of j according to Eq. (4). Obviously we would hardly think of classifying, e.g. animals in a population with the resolution of an atom but molecules, e.g. DNA are routinely sequenced to this precision. For example the gene length distribution of *Escherichia coli* is also found skewed [19].

To qualify an entity to the distribution depends often also on its functional properties. For example, a seemingly simple variable ‘the weight of a female wolf’ is in fact a complex criterion. Variation in weights results from fluctuations in a number of nested hierarchical steps of assembly and each step is subject to some criteria to yield the characteristic functional properties.

According to Eq. (4) the density-in-energy, i.e. the amount of matter and energy in a class $j + n$, is

$$\phi_{j+n} = \exp[(j+n)\phi_1] = \phi_j \exp(n \ln \phi_1). \quad (5)$$

A distribution of densities-in-energy $\phi_{j-n \dots j+n}$ in adjacent classes $j - n \dots j + n$ is obtained in the familiar logarithmic form

$$\mu_{j-n \dots j+n} = RT(\ln \phi_j + \sum_n n \ln \phi_1) = \mu_j + \sum_n \mu_n, \quad (6)$$

which is normal according to the central limit theorem, when $n \ll j$. Thus the distribution of adjacent classes resulting from a natural process is log-normal. The mean μ_j corresponds to the most probable parameter value. Thus we conclude that the principle of increasing entropy underlies log-normal distributions that are the stationary-state distributions of natural processes.

The spread and skew of a distribution depends on how much the non-linear density-in-energy changes within the span of variation that is typically determined by functional criteria (Fig. 2). For example, without the functional factor of a female wolf, weights x_j of all kinds of objects with densities-of-states ϕ_j in all kinds of classes j would qualify to the distribution. When the functional criteria of energy transduction are acknowledged the distribution begins at the class j_0 that meets

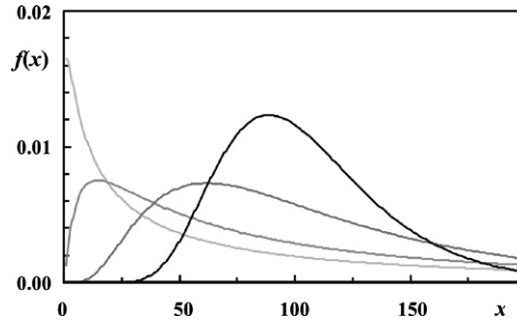


Fig. 2. Skewed distributions result from natural processes when the system has evolved toward more probable states. The states are customarily described in terms of thermodynamics by exponential densities-in-energy. The amount of variation ($\ln\sigma = 1.4, 2.0, 4.0, 8.0$), in the densities-in-energy about the mean ($\ln x = 100$) dictate the spread and skew of the distribution.

the minimum qualifications (Fig. 2). Thus the parameter value x_{j0} associates with ϕ_{j0} . It is the natural unit of normalization [11] for the dimensionless parameter distribution $x = x_j/x_{j0} \geq 1$. The positive lower bound $x = 1$ coincides with the minimum functional characteristics associated with the parameter. In this special case a failure in any step of assembly would lead to a disqualification.

Often a biological assembly process will terminate naturally because a functionally deficient entity cannot sustain subsequent jk -transitions of energy transduction that would be vital for further growth, differentiation or in general evolution. Spontaneous degradation processes, exoergic reactions are usually not subject to particularly stringent functional criteria. When the natural process requires no sophisticated mechanisms to advance the distribution comes with a wide variation. For example a grain size distribution is wide with a substantial skew because the parameter does not carry much functional specifications about the grinding process.

In general natural distributions speak about flows of matter and energy down along the free energy gradients during natural processes. The flows are subject to non-linear thermodynamic confines expressed by chemical potentials. These are encountered in stochastic search for more probable states. An unobstructed abstract process, i.e. one not involving matter and energy, will give a symmetric distribution, i.e. the Gaussian distribution. An obstructed ideal process, e.g. a random flight next to a wall will yield a skewed, however not log-normal, distribution [20,21].

4. Time courses towards stationary states

Many growth processes yield cumulative curves. When plotted at logarithmic scale the growth curve is a sigmoid of time corresponding to the cumulative curve of a skewed distribution. Also dose–response curves are of this nature [5,12].

The entropy increase rate [15] of a system is obtained from the time derivative of Eq. (2)

$$\frac{d \ln P}{dt} \approx \sum_{j=1} \frac{dN_j}{dt} \ln \left(\frac{\prod_k \phi_k}{\phi_j} \right) = \frac{1}{RT} \sum_{j=1} v_j \left(\sum_k \mu_k + \Delta Q_{jk} - \mu_k \right) = \frac{1}{RT} \sum_{j=1} v_j A_j, \quad (7)$$

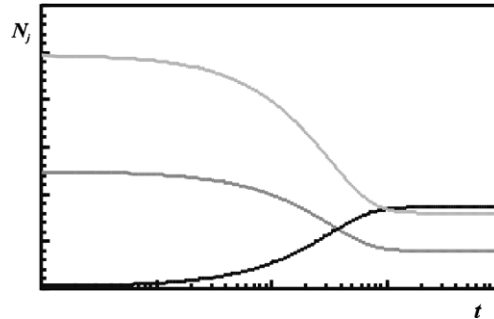


Fig. 3. Cumulative concentrations curves N_j vs. time t obtained from a simulation of an example chemical reaction $2N_1 + N_2 \leftrightarrow N_3$ using Eq. (8). The substrates (1 and 2) are consumed when the product (3) is forming until the stationary state given by Eq. (3) is reached. The sigmoid curve can be closely approximated, e.g. by the logistic equation with an initial growth rate r and a constant carrying capacity K . However, the curves as well as Eqs. (7) and (8) reveal that the evolution rate is following changing thermodynamic forces, the diminishing free energy differences between the substrate and product densities-in-energy.

where the flow $v_j = dN_j/dt$ of matter to entities N_j is driven by the potential difference, thermodynamic driving force also known as affinity [14] A_j , towards the equilibrium where $A_j = 0$. The conservation of matter requires that a flow v_j to N_j is exactly opposite to v_k from N_k consistently with stoichiometries of reactions [15]

$$v_j = r \ln \frac{\prod_k \phi_k}{\phi_j} = r \frac{A_j}{RT}, \quad (8)$$

where r depends on particular reaction or activation mechanisms. The cumulative curves of $N_j(t)$ are sigmoids (Fig. 3). These time courses are often approximated by logistic equations or modeled as stochastic processes and analyzed using mathematical and statistical techniques [4].

Likewise, a sigmoid action–response v_j vs. $\ln N_k$ curve arises when the response rate v_j is limited by the dissociation rate r of activator–product complex N_* . At this steady state, flows to and from an activated complex vanishes [22]

$$v_* = r_a \ln(\phi_k \phi_a / \phi_*) - r \ln(\phi_a \phi_j / \phi_*) = 0, \quad (9)$$

where $RT \ln \phi_*$ denotes the chemical potential of the activated complex of the substrate k and the activator a , e.g. catalyst. When $r_a \gg r$ the potential of dose $\ln \phi_* = \ln(\phi_k \phi_a)$ limits v_j (Fig. 4). The sigmoid cumulative curves from natural processes have no analytical expression because the motion of probability is intrinsically non-integrable in the sense of Poincaré [15]. Motion is driven by the potential energy differences that keep changing during the evolution. Thus there is no transformation that would separate the driving forces from the probability in motion. This is an inherent property of dissipative evolution.

Natural processes such diffusion, spontaneous and autocatalytic reactions, have been related to the log-normal distributions [3,11]. The typical form is $\ln x_j(t) = \ln x_{j0} + f(t)$ where the function $f(t)$ is subject to normally distributed fluctuations about the most probable value. It is obtained from integration of dN_j/dt . For example a process $dN_j/dt = -rN_j$ will give $f(t) = -rt$ where r is the first order-rate constant subject to fluctuations. Likewise an autocatalysis $dN_j/dt = rN_j$ will give $f(t) = rt$. The exponential, i.e. Malthusian form, accounts well for the initial growth phase that

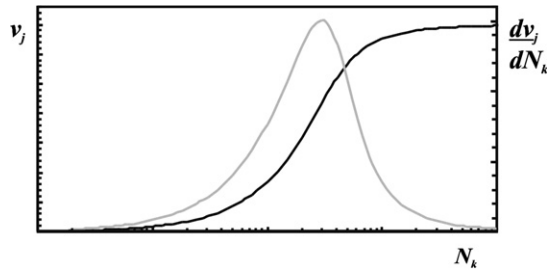


Fig. 4. Response rate v_j vs. a dose concentration N_k obtained from simulations of Eq. (9). The derivative curve (grey) peaks at the dose concentration corresponding to half-maximal affinity when half of the initial thermodynamic driving force has been exhausted.

is often followed by power-law dependence. The logistic form or Verhulst–Pearl equation by its two-timing form [23] approximates well courses due to diminishing potential energy differences.

5. Discussion

The principle of increasing entropy by decreasing free energy underlies many processes in nature. We find that the resulting parameter distribution characterizing matter and energy in aggregates is skewed because densities-in-energy, i.e. chemical potentials of entities are non-linear functions. When the natural process is subject to fluctuations a log-normal distribution closely captures the resulting natural distribution. This finding is consistent with earlier results. Chemical reactions, autocatalytic processes and diffusion as a function of time, all natural processes, have been associated with log-normal distributions consistent with our reasoning about their common basis [11,12]. Skewed distributions have also emerged when outcomes of historical and evolutionary sequences have been placed in categories of time as demonstrated here for a simple model system undergoing chemical reactions.

The principle of increasing entropy directs the flows of matter and energy toward more probable partitions along the fast routes of entropy increase [15]. The existence of a preferred path, i.e. the steepest gradient, clarifies the validity of small fluctuation approximation, i.e. variation in the chemical potential can be linearly approximated when $\ln(1 + \delta N_j) \approx \delta N_j$ and $\exp(\delta \epsilon) \approx 1 + \delta \epsilon$. The meaning of linearity is more profound than a mere approximation – it often defines entities that are qualified to a distribution. An entity that contains substantially more or significantly less matter and energy than its counterparts may not be regarded as an entity that should belong to the parameter distribution.

The natural distribution is narrow and the skew is small when entity assembly expenses in terms of matter and energy are large and accompanied functional criteria are stringent. Then fluctuations are inherently small since a large fluctuation would be either thermodynamically expensive or cause a functional failure in the continuation of the natural process. In this case the resulting distribution may even be mistaken for an additive normal distribution because the underlying chemical potential does not change markedly over the parameter span. For example, parameter distributions of animal populations are often narrow and without much of a skew. On the other hand the natural distribution is wide and its skew is large when the classification criteria allow to

include results from a wide spread of activities. In these cases large fluctuations are thermodynamically afforded and they do not jeopardize the continuation of process. For example, a distribution of incomes contains both minimal individual efforts required for survival and multinational orchestrated enterprises granting superfluous existence.

The entropy expression by Eq. (2) not only enumerates entities but includes also potential energy gradients and identifies interactions, i.e. mechanisms of energy transduction that drive evolution. The interactions define an entity, e.g. a biological unit, by properties, i.e. phenotype that support its existence. Thus the natural statistics denotes all the attributes that we may recognize when accepting an entity to a distribution.

It may appear amazing that a seemingly simple formula of entropy that only serves to express probable scenarios and outcomes accounts for copious complex phenomena. The statistical description in the multiplicative form remains the same despite increasing complexity because the equation for S is self-similar for various scales of matter, energy and time. Following the reasoning by Jacobus C. Kapteyn, we conclude that the connection between the log-normal form of the curve and its cause is the principle of increasing entropy. It leads us to the precious information about the nature of many processes.

Acknowledgments

Discussions with Martti Louhivuori, Michael Nilges and Vivek Sharma have been most enlightening.

References

- [1] D. McAlister, Proc. R. Soc. 29 (1879) 367.
- [2] J.C. Kapteyn, *Skew Frequency Curves in Biology and Statistics*, Astronomical Laboratory, Noordhoff, Groningen, 1903.
- [3] J.H. Gaddum, Lognormal distributions, *Nature* 156 (1945) 463.
- [4] C.J. Mode, C.K. Sleeman, *Stochastic Processes in Epidemiology: HIV/AIDS, Other Infectious Diseases and Computers*, World Scientific Publishing, 2000.
- [5] E. Limpert, W.A. Stahel, M. Abbt, Log-normal distributions across the sciences: keys and clues, *Bioscience* 51 (2001) 341.
- [6] J. Heitzenberg, Properties of log-normal particle size distribution, *Aerosol Sci. Technol.* 21 (1994) 46.
- [7] J.M. Mäkelä, I.K. Koponen, P. Aalto, M. Kulmala, One-year data of submicron size modes of tropospheric background aerosol in southern Finland, *J. Aerosol Sci.* 31 (1999) 595.
- [8] J. Aitchison, J.A.C. Brown, *The Log-normal Distribution*, Cambridge University, Cambridge, 1957.
- [9] E.L. Crow, K. Shimizu (Eds.), *Log-normal Distributions: Theory and Applications*, Dekker, New York, 1988.
- [10] C. Gut, E. Limpert, H. Hinterberger, A computer simulation on the web to visualize the genesis of normal and log-normal distributions. <http://stat.ethz.ch/vis/log-normal>, 2000.
- [11] A.L. Koch, The logarithm in biology I. Mechanisms generating the log-normal distribution exactly, *J. Theor. Biol.* 12 (1966) 276.
- [12] A.L. Koch, The logarithm in biology II. Distributions simulating the log-normal, *J. Theor. Biol.* 23 (1969) 251.
- [13] M. Alonso, E.J. Finn, *Fundamental University Physics Quantum and Statistical Physics*, vol. 3, Addison-Wesley, London, 1983.
- [14] D. Kondepudi, I. Prigogine, *Modern Thermodynamics*, Wiley, New York, 1998.

- [15] V. Sharma, A. Annala, Natural Process – Natural Selection, *Biophys. Chem.* 127 (2007) 123.
- [16] J.W. Gibbs, *The Scientific Papers of J. Willard Gibbs*, 1. OxBow, Woodbridge, CT, 1993–1994.
- [17] P.W. Atkins, *Physical Chemistry*, Oxford University, New York, 1998.
- [18] P. Waage, C.M. Guldberg, *Forhandlinger*, 35. Videnskabs-Selskabet i Christiana, 1864.
- [19] M. Skovgaard, L.J. Jensen, S. Brunak, D. Ussery, A. Krogh, On the total number of genes and their length distribution in complete microbial genomes, *Trends Genet.* 17 (2001) 425.
- [20] S. Chandrasekhar, *Rev. Mod. Phys.* 15 (1943) 1.
- [21] M. Louhivuori, K. Pääkkönen, K. Fredriksson, P. Permi, J. Lounila, A. Annala, On the origin of residual dipolar couplings from denatured proteins, *J. Am. Chem. Soc.* 125 (2003) 15647.
- [22] D.L. Nelson, M.M. Cox, *Lehninger Principles of Biochemistry*, third ed., Worth Publishers, USA, 2000.
- [23] S. Strogatz, *Non-linear Dynamics and Chaos*, Perseus Books, Cambridge MA, 1994.