GAUSSIAN/NON-GAUSSIAN DISTRIBUTIONS AND THE IDENTIFICATION OF TERRESTRIAL AND EXTRATERRESTRIAL INTELLIGENCE OBJECTS

SERGEY HAITUN — Candidate of Physics and Mathematics, Institute for the History of Science and Technology (Moscow, Russia)

E-mail: haitunsd@mail.ru

Statistical criteria used today in the analysis of radio signals suspected on reasonable extraterrestrial origin, are based on the assumption that all the radio signals of natural origin are described by a Gaussian distribution, which is traditionally understood as the Gauss distribution. Usually the normal (Gauss) distribution is opposed to all the others. However, this is difficult to recognize the reasonable, because in nature there are many different distributions. The article offers a more realistic dichotomy: the Gaussian distributions, obeying the central limiting theorem, dominate in nature, while non-Gaussian ones, obeying the Gnedenko-Doeblin limiting theorem, are generated by intelligent beings. When identifying objects belonging to an extraterrestrial civilization described by a non-Gaussian distribution is preferable to use the rank form distributions. Using this criterion is associated with certain difficulties: (1) in nature there are also non-Gaussian distributions; (2) in their activities animals generate non-Gaussian distributions like humans; (3) the identification of non-Gaussian distributions in the rank form is hampered sometimes by the rank distortion effect of mathematical nature.

Key Words: SETI, Gaussian distribution, non-Gaussian distribution, the central limiting theorem, the Gnedenko-Doeblin limiting theorem, the Zipf distribution, Zipfian distribution, frequency distribution, rank form distribution, the rank distortion effect, ants.

Introduction

The beginning of the Global Project SETI (Search for Extra-Terrestrial Intelligence) is sometimes connected with publication of the article [Cocconi, Morrison, 1959]. If you rely on this date, the SETI is more than 55 years old, but, as we know, the positive results have not been obtained so far.

A common view is that we do not know whether there is an intelligent life in the Universe. Personally, I think [Haitun, 2005: p. 280–285; Haitun, 2006: p. 210–215], that due to infinity of the Universe, an intelligent life beyond the Earth exists definitely, and that, moreover, the number of its centers is infinite. The question is: how far these centers are spaced from each other? There are two opportunities: if the Earth noosphere is the only one in our Metagalaxy, then we are doomed to cosmic solitude, but if there are millions similar noospheres in our Metagalaxy, then our chances of meeting with colleagues in the mind are quite real.

Although we do not know today what is the real situation, we should be ready for such a meeting, as if we knew for sure that this meeting awaits us (future planning should be alternative). Among other things, we should learn how to recognize extraterrestrial civilizations. So in general terms the SETI strategy seems quite correct. However, there are certain claims to the SETI means (tools) of recognition.

© Sergey Haitun, 2015
According to the well-known monograph [Gindilis, 2004], scientists working in the SETI, in many respects rely on search of meaningful radio signals from the space. This, of course, is reasonable, and the search for meaning in the space signals can be successful if these signals will be certainly meaningful (meaningful to a large extent), i.e. if, say, we will receive a sequence of prime numbers, or something like that. But one cannot rely only on this scenario. What if the signals are only partly meaningful, i.e. they are partly meaningful and partly meaningless? It could be possible that an extraterrestrial civilization, from which the signal comes, has entirely different meaning than we have. The concept of meaning has the flaw that it is irreproducible, for one researcher this signal may seem sensible, and for another it is senseless.

Thus, we come to a conclusion, that we need formal criteria which allow us to identify natural objects which have arisen without the mind, and artificial objects without appealing to their meaningfulness or meaninglessness.

Such are the statistical criteria, also used in the SETI, but it is not all right with them. The fact is that, according to [Gindilis, 2004], using a statistical criterion in the framework of the SETI is based on the dichotomy: Gaussian/non-Gaussian distribution, which appears to be correct. However, a Gaussian distribution is understood simply as the normal distribution, and under a non-Gaussian distribution (without mentioning that) is understood any distribution other than the normal one. The normal distribution associated with the objects of natural origin, is opposed to all other distributions. This is difficult to accept when we distinguish natural objects from artificial objects, because in nature there are many distributions other than the normal one.

In my opinion, the other, more general dichotomy: Gaussian/non-Gaussian distributions, is more rational. These are two large classes of distributions. Gaussian distributions are subject to the central limiting theorem, and non-Gaussian ones obey the Gnedenko-Doeblin limiting theorem. I would argue that the first dominate in nature, while artificial objects (created by the mind) are described by distributions of the second class [Haitun, 1982; Haitun, 1983; Haitun, 1989]. It is suggested to determine objects of the earthly or unearthly nature in accordance with distributions describing them (Gaussian or non-Gaussian).

As you can see, the suggested idea sounds simple enough. However, its implementation requires a mastery of the relevant apparatus.

1. Gaussian/non-Gaussian probability distributions

The dichotomy: Gaussian/non-Gaussian distributions has a different meaning depending on about which distributions we talk, because we could talk about probability distributions, built on samples from infinite volume, or about statistical distributions, built on samples from a final volume.

The modern theory of probability and mathematical statistics are built on the limiting theorems on convergence of distributions of sums of identically distributed random independent variables with increasing sample size to the so-called stable distributions. We divide distributions converging to stable ones into two classes — Gaussian and non-Gaussian distributions. Distributions belonging to the first class obey the central limiting theorem, while those belonging to the second class obey the Gnedenko-Doeblin limiting theorem [Gnedenko, 1939; Doeblin, 1940].

It is easy to understand why the limiting theorems on convergence of distributions to stable distributions are so important for scientific studies. If a distribution of
measured objects does not converge to a stable one with increasing sample size, it means that this distribution does not have a specific form, i.e. its form depends on a sample size. In this case the parameters of such a distribution (the mean, median, dispersion, quantile et al.) which are determined by its form do not have certain values. Such a measurement is not reproducible, that is not allowed by a scientific research. We (scientists) restrict ourselves by reproducible measurements and therefore we work only with distributions obeying the central limiting theorem or the Gnedenko-Doeblin limiting theorem. So any distribution which is really interesting to science can be either Gaussian or non-Gaussian distribution, and no middle ground.

The central limiting theorem applies to distributions converging in the above sense to the Gauss distribution, i.e. to the normal distribution. According to this theorem, a necessary and sufficient condition for a distribution to converge to the Gauss distribution is that the mean and dispersion of such a distribution be finite. Such distributions we call Gaussian. Note that the theorem says nothing about the form of Gaussian distributions.

I would like to say a few words about terminology. In the scientific literature, the terms “Gaussian distribution” and “the Gauss distribution” are not distinguished (following the language rules). Usually they both refer to the same distribution, which is also called the normal one. For our goal, however, it is important to distinguish them. Therefore, we use the term a Gaussian distribution in the case that this distribution obeys the central limiting theorem (and there are a lot of such distributions), while the Gauss distribution is the one and only one which is also called the normal distribution. The Gauss distribution is one of Gaussian distributions, but a Gaussian distribution, in the general case, is not the Gauss distribution. In other words, we prefer to break a bit language rules to ensure the required accuracy of terminology.

The Gnedenko-Doeblin limiting theorem applies to a distribution, not converging (in the sense described above) to the Gauss distribution, but converging to another stable distribution. For this it is necessary and sufficient that the distribution \( f(x) \) for large values of the random variable \( x \) has the form of the Zipf distribution (up to a slowly varying function)

\[
n(x) = \frac{C}{x^{1+\alpha}}, \quad 0 < x_0 \leq x \leq 0 < \alpha < \infty, \quad (1)
\]

with \( \alpha < 2 \). Here \( n(x) \) is a frequency; \( \alpha \) is the Zipf distribution characteristic exponent; the parameter \( C \) provides the distribution normalization, i.e. \( \sum n(x) = N \) where \( N \) is the sample size.

A distribution which for large values of the variable has a form of the Zipf distribution with any \( \alpha \) we call a Zipfian distribution.

I would like to say a few words about terminology again. In the literature the terms “a Zipfian distribution” and “the Zipf distribution” do not differ. They both are used for the same distribution given by Eq. 1. However, for our goal we need to distinguish between them. A Zipfian distribution is a distribution which has
the form of the Zipf distribution (there are many such distributions) at large values of the variable, whereas the Zipf distribution is the only one distribution given by Eq. (1). The Zipf distribution belongs to the Zipfian distributions, however a Zipfian distribution, in the general case, is not the Zipf distribution. In other words, we prefer again to break a bit language rules to ensure the required accuracy of terminology.

In the logarithmic coordinates a Zipfian distribution at large values of the variable \( x \) is inclined to the horizontal axis at an acute angle whose tangent determines the Zipf distribution characteristic exponent \( \alpha \)

\[
\lim_{x \to \infty} \frac{d \ln n(x)}{d \ln x} = -(1 + \alpha),
\]

(2)

(see Fig. 1).

Fig. 1. Examples of Zipfian frequency distributions in logarithmic coordinates. The curve 1 is the generalized Pareto distribution \( n(x) = \left( \frac{\alpha}{x^{\alpha}} \right) \left[ \left( \frac{x}{\alpha} \right)^{\alpha-1} \right], x \geq 0 \). The curve 2 is the Cauchy distribution \( n(x) = \frac{2N}{\pi d^2} \left[ 1 + \left( \frac{x}{d} \right)^2 \right], x \geq 0 \). The curve 3 is the Zipf distribution. In these coordinates Zipfian distributions for \( x \to \infty \) have asymptote inclined to the horizontal axis at an acute angle whose tangent determines the Zipf distribution characteristic exponent \( \alpha \).

It is not difficult to see that for the Zipf distribution moments of the order \( n \) are infinite for \( \alpha \leq n \), i.e. the mean, or the first-order moment, is infinite for \( \alpha \leq 1 \), the dispersion or the moment of the second order, is infinite for \( \alpha \leq 2 \). The same is true for Zipfian distributions. For instance, at \( \alpha \leq 1 \) the mean and dispersion are infinite, and at \( 1 < \alpha \leq 2 \) the mean is finite, while the dispersion is infinite; at \( \alpha > 2 \) the mean and the dispersion are both finite.

Thus, we can say that (up to slowly varying functions mentioned in the Gnedenko-Doeblin limiting theorem) a non-Gaussian probability distribution is a Zipfian distribution at \( \alpha \leq 2 \) (although at \( 1 < \alpha < 2 \) the mean is finite), while a Zipfian distribution is a Gaussian distribution at \( \alpha > 2 \).
Gaussian non-Zipfian distributions (Gauss, Poisson, and others) can be also presented in the logarithmic coordinates. For these distributions

$$\lim_{x \to \infty} \frac{d \ln n(x)}{d \ln x} = -\infty.$$  \hfill (3)

Comparing with Eq. (2), we see that

$$\alpha = \infty.$$  \hfill (4)

In other words, in these coordinates a Gaussian non-Zipfian distribution has the asymptote (at $x \to \infty$) inclined to the horizontal axis at the right angle (see. Fig. 2).

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig2.png}
\caption{Examples of Gaussian non-Zipfian frequency distributions in the logarithmic coordinates. The curve 1 is the “right” Gauss distribution $f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{x^2}{2\sigma^2}\right]$, $x \geq 0$. The curve 2 is the Poisson distribution $f(x) = \frac{\exp(-\lambda)\lambda^x}{x!}$, $x \geq 0$. The Curve 3 is the lognormal distribution $f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)^2\right]$, $x \geq 0$. In these coordinates Gaussian non-Zipfian distributions are tangent inclined to the horizontal axis at right angles at $x \to \infty$.}
\end{figure}

\section{2. Gaussian / non-Gaussian statistical distributions}

The finite/infinite dispersion and the threshold value of the parameter $\alpha$ serve criteria to distinguish between the Gaussian and non-Gaussian distributions only in the case of probability distributions. Meanwhile, a probability distribution is nothing more than a mathematical abstraction, which assumes an infinite general population from which samples of infinite size are taken. In contrast, in this paper we are mainly interested in empirical distributions which are based on finite-size
samples taken from finite general populations. It is clear that the dispersion and other moments of distributions based on such finite general populations and finite samples are themselves finite. In this situation, the dichotomy of Gaussian/non-Gaussian distributions becomes the dichotomy of Gaussian/non-Gaussian general populations.

*If for a given general population the dispersion increases significantly with the sample size, then this general population we call non-Gaussian, in the opposite case we call it Gaussian. The less the parameter $\alpha$, the more substantial this dependence.*

Thus, for statistical distributions the transition from Gaussian distributions to non-Gaussian ones occurs continuously: for small parameters $\alpha$ (say, for $\alpha$ about 3 or less) Zipfian distributions are non-Gaussian and for large $\alpha$ (say, for $\alpha$ greater than 10 or 15) they are Gaussian. For intermediate values of $\alpha$ (for $3<\alpha<10$ or for $3<\alpha<15$) we have an intermediate situation.

We see that the Zipfian distribution characteristic exponent $\alpha$ plays the central role in determining the character of distributions (Gaussian/non-Gaussian). The form of a distribution of sample values of any distribution statistic (including $\alpha$), is determined, as is known, by the method chosen to calculate sample values of this statistic. If the maximum likelihood method is chosen to determine the parameter $\alpha$, then the distribution of the $\alpha$ sample values is asymptotically normal. In the case of the Zipf distribution

$$\sigma_{\alpha} = \frac{\alpha}{\sqrt{N}},$$

(5)

where $N$ is the sample size [Kramer, 1948: p. 547].

Not being a professional mathematician, first, I cannot say how sample values of $\alpha$ are distributed and second, what is the error of these values when using other methods of determining them, for example, when we use the graphic method together with rank form distributions (see next section). It seems, however, that the estimate (5) overestimates the error calculated by the graphic method.

### 3. Rank form of statistical distributions

When working with non-Gaussian distributions is often fruitful to employ rank form distributions, which, from my point of view, is the most convenient for using the statistical criterion discussed in this article.

The rank form exists for all statistical distributions, i.e. for all discrete distributions built on samples of finite size. The advantage of the rank form distributions compared to frequency distributions is due to the possibility to work with small sample sizes, say 3, 4, 5. The statistical accuracy required here is achieved by increasing the accuracy of measurement values of the random variable. For frequency form distributions sufficient statistical reliability is ensured only by a large sample size, which is practically not always possible.

A frequency form distribution determined by the frequency $n(x)$ with which a given value of the random variable $x$ occurs in a sample of the volume $N$: 
\[ \sum_{x=x_{\min}}^{x=x_{\max}} n(x) = N \]  

where \( x \) and \( J \) are the minimum and maximum sample values of \( x \).

**A rank form distribution** is determined by

\[ r(x) = \sum_{\xi=x}^{J} n(\xi). \]

The rank \( r \) means a sequence number of a given value of the random variable \( x \), when these values are arranged in order of decreasing \( x \). For a sample of size \( N \), we have \( N \) ranks. **Equal values of** \( x \) **have different ranks.**

**The frequency differential form of the Zipf distribution** is determined by Eq. (1).

The frequency integral form of the Zipf distribution is given by

\[ F(x) = \frac{C}{\alpha N} \left( \frac{1}{x_0^\alpha} - \frac{1}{x^\alpha} \right) = 1 - \frac{x_0^\alpha}{x^\alpha}. \]

The rank differential form of the Zipf distribution is given by

\[ x(r) = \frac{A}{(r+B)^\gamma}, \quad \gamma = \frac{1}{\alpha}, \quad A = \left( \frac{N-1}{1/x_0^\alpha - 1/J^\alpha} \right)^{1/\alpha}, \quad B = \frac{N-1}{(J/x_0)^\alpha - 1}. \]

In logarithmic coordinates a Zipfian rank distribution is a straight line, inclined to the axes at an acute angle at large values of the random variable \( x \) (i.e. for ranks 1, 2, 3, ...) (see Fig. 3).

**Fig. 3.** Examples of Zipfian distributions in the rank form (in logarithmic coordinates). The curve 1 is the generalized Pareto distribution, the curve 2 is the Cauchy distribution, the curve 3 is the Zipf distribution. At \( x \to \infty \) the rank Zipfian distributions have asymptote inclined to the vertical axis at an acute angle. The value of this angle determines the Zipfian distribution characteristic exponent \( \alpha \). For clarity, we consider the case when the rank distortion effect (see. Sec. 10) is absent.
4. Non-Gaussian nature of social phenomena

Social phenomena are non-Gaussian. The author established this fact working with a random sample of 190 social stationary (independent of time) distributions. Figures 4 and 5 show 18 of such distributions, Fig. 6 shows the resulting distributions.

Fig. 4. Stationary scientometric distributions which cannot be approximated by the Zipf distribution (according to data of literature sources): 5.1. Rank distribution of periodicals by their citation. The Gauss distribution with the same sample size and mean and mean quadratic deviation is presented for comparison; 5.2. Rank distribution of countries by the number of scientific journals; 5.3. Distribution of the number of authorships by the number of authors for an article; 5.4. Rank distribution of periodicals by the number of annual library requests; 5.5. Rank distribution of scientists by partial productivity (for an article with n authors each receives the mark 1/n); 5.6. Rank distribution of 152 journals from the whole world, whose publications of 1967–1968 were most often referred to in 1969; 5.7. Rank distribution of nine subject fields by the number of references per 100 pages of periodicals; 5.8. Rank distribution of the names of scientists by the frequency of occurrence in eight textbooks of psychology; 5.9. Distribution of publications by the number of authors. Coordinates are logarithmic.
Haitun Sergey. Gaussian/non-Gaussian distributions and the identification of terrestrial and extraterrestrial intelligence objects

Fig. 5. Stationary non-scientometric distributions of human activity (according to data of literature sources): 6.1. Rank distribution of the Russian alphabet letters by the frequency of occurrence; 6.2. Rank distribution of countries by the capita electricity production; 6.3. Rank distribution of Soviet Republics by cement production in the USSR in 1975; 6.4. Rank distribution of 27 enterprises of British electrical engineering industry by unit out of production in 1933–1934; 6.5. Wage distribution of population; 6.6. Distribution of the USA population by private income; 6.7. Distribution by annual income of persons in the United Kingdoms surtaxed during a year beginning from April 6, 1950; 6.8. Rank distribution of nationalities in the USSR by their numbers in 1970; 6.9. Cumulative frequency distribution of billiard scores in 50 services. Coordinates are logarithmic.
It turns out that it is possible to approximate all considered social distributions by the Zipfian distributions with predominantly small values of the parameter $\alpha$ and only in 6 cases from 190 we have obtained $\alpha > 10$. The distribution of the values of the parameter $\alpha$ itself is also the Zipfian distribution with $\alpha \approx 1.31$. All this allows us to talk about non-Gaussian nature of social phenomena, actually about domination of non-Gaussian distributions among stationary distributions describing reasonable activities.

5. Clarification 1: non-stationary distributions are not considered

Social distributions can be divided into two categories: stationary distributions not involving time dependence and non-stationary ones which contain the time. It turns out that social phenomena described by non-stationary distributions have non-Gaussian nature, while, as shown by empirical data, social phenomena described by stationary distributions can have different forms of Gaussian and non-Gaussian nature as well. We have established the phenomenon of non-Gaussian nature of social phenomena only after having identified social stationary distributions in a special class.

Naturally that the radio signals coming from the space, analyzed for their “reasonableness”, are distributed in time. This does not mean, however, that they cannot be analyzed using the criterion of Gaussian/non-Gaussian distributions. Of course, it is possible, but in this case the time cannot be the random variable with
which a distribution is built. For example, we have a radio signal recorded during the time interval $T$. We have the right to consider this signal as a set (a general population) of values of signal amplitude, without associating these values with time. Then we build our distributions on this general population taking from it random samples of amplitude values. The resulting distributions, rank or frequency ones, will be stationary as required.

6. Clarification 2: We do not take into account distributions constructed with using closed scales

When establishing that social phenomena have non-Gaussian nature we have also deliberately excluded empirical social stationary distributions constructed with help of points, bounded above by the measurement procedure, i.e. with help of the closed scales. All such distributions are Gaussian distributions, for example, the distribution of individuals by the IQ (see. Fig. 7).

Fig. 7. The distributions of individuals by the IQ. The linear coordinates. It is often approximated by the Poisson distribution.

The distribution of individuals by the IQ has a short tail and is approximated by a usual Gaussian non-Zipfian distributions including the Poisson distribution. How does this consist with the thesis that social distributions are non-Gaussian?

The point is that the closed scales are incorrect in principle. They generate non-additive variables. Each of these variables forms a set of its values on which the operation of addition does not work.

At the same time the operation of addition is the basis of quantitative mathematics (the quantitative mathematical operations — subtraction, multiplication, division, differentiation, integration, etc. — are based on addition). It is considered in details in the monograph [Haitun, 1989]. Here only the main points of our analysis are given.
Today the perception of non-additivity of a variable as an objective property, i.e. independent on a subject of measurement, is common to all researchers. To shake this belief, we will pay attention to the following observation: non-additivity arises when short-tailed distributions are used.

Let us illustrate this with the example of the scientific output of a researcher which is measured often by a number of citations. The distribution of scientists on the number of citations is non-Gaussian (long-tailed), while the measurement result is additive. We can sum the numbers of citations. You can, however, ask experts to assess the scientific output of researches, say, on the five-point scale. In this case the distribution of researchers will be short-tailed (Gaussian), while the measurement result is not additive because there is no sense to sum the received points. However, in these two cases we have measured the same value — the scientific output of a researcher. Thus, we can see that non-additivity is generated by the measurement procedure.

The same would happen if we, for example, ask some experts to estimate the mass or weight of physical bodies using the five-point scale. The results of such measurements would also be non-additive. The point is certainly not in the variable “weight” or “mass”, it is and in the procedure of expert measurement.

In the measurement theory the variables “weight”, “mass” or, say, a “scientific output of a researcher” are called latent variables or shorter latents (introduced by the author), i.e. directly unobservable variables, which is reflected in their name: latent (i.e. this variable is not possible to measure directly). Therefore, the values of quantities, directly observable, called the indicators, are always measured. But their values characterize the latent values indirectly. As indicators can be the number of citations, the expert score, the displacement of the arrow of the spring balance, etc. It looks as if indicators and latents live in different planes, never crossing.

Obviously, the properties of the indicator must reproduce the latent properties. In particular, the form of indicator distribution (i.e. the distribution of indicator values built with the sample of measurement objects) must coincide with the form of latent distribution. If a latent is additive, than non-coincidence of forms of the indicator distribution and the latent distribution makes the indicator non-additive. This is what happens when you use closed scales which deform the indicator scale with respect to the latent scale in the region of large values of the indicator values (see. Fig. 8).

To eliminate the indicator deformation relative to the latent we should open closed indicator scales. But then most social stationary distributions built with their help will become long-tailed (non-Gaussian). In saying this, we rely on the fact that such distributions are obtained when the open indicator scale is used for the social latent.

7. How to apply the discussed statistical criterion

First, we are building the rank distribution of the analyzed object of unknown origin in logarithmic coordinates, second, we are approximating the empirical points by the straight line (if it is possible), or are building the asymptote for large values of the random variable x (that is always possible), third, we are looking for the sample value of the parameter α by the tangent of the angle of inclination of the straight line or by the asymptote to the ordinate axis (or by the cotangent of the
angle of inclination to the horizontal axis). If $\alpha$ is small, say, $\alpha < 3$, we believe that this object has a reasonable origin, if $\alpha$ is large, say, $\alpha > 10$ or $\alpha > 15$, the object has a natural origin, and for intermediate values of $\alpha$, $3 < \alpha < 10$ or $3 < \alpha < 15$, (it is our choice), the nature of the object is unknown to us.

Let us consider two examples of calculation of the two real objects.

The window 1 in my apartment. There are three main window parameters: 151 cm (height), 87 cm (width of the main window sash) and 37 cm (width of the window leaf). We arrange these parameters in the descending order (151, 87 and 37) and assign rank values to them 1, 2, 3. Then we put the rank values on the horizontal axis, and the window parameters on the ordinate axis. Thus we have received three points on the graph (see. Fig. 9).
The window 2 in my apartment is larger than the first one, so it has four basic parameters: 151 cm (the height), 71 cm, 71 cm (two window sashes have the same width), and 37 cm (the window leaf width). The corresponding ranks are 1, 2, 3 and 4. The four points of the graph for this window also lie on a straight line with a small spread around it. But the angle of inclination of this line to the horizontal-axis is slightly smaller than the previous one (see. Fig. 10).

![Figure 10](image)

**Fig. 10.** The window 2. The logarithmic coordinates.

Small values of $\alpha$ confirm that the windows in my apartment are of human (reasonable) origin, for sure they are not of natural origin. So that the demonstrated examples speak for the use of the discussed statistical criterion.

8. **the first pitfall: in nature there are also non-Gaussian distribution**

There are Zipfian distributions with small values of the parameter $\alpha$ in the inorganic and organic worlds. For instance, distributions of anions and cations by the occurrence in seawater have such form. Having used the data contained in the book by Yu. Odum [Odum, 1986: p. 286], we have built the rank distribution of 5 anions according to their number in 1 kg of seawater (the Zipfian distribution with $\alpha \approx 0.19$) and the similar distribution of 4 cations (the Zipfian distribution with $\alpha \approx 0.37$).

Distributions of chemical elements according to their occurrence in the earth crust and in our Metagalaxy are also non-Gaussian. Using the table in the book [Folsom, 1982: p. 16], we have built the rank distribution of 15 chemical elements according to their prevalence in our Metagalaxy (the Zipfian distribution with $\alpha \approx 1.18$).

Distributions of cosmic bodies by mass [Trubnikov, 1993, 1995] and cosmic particles by energy are also non-Gaussian. The distribution of earthquakes by
energy is the Zipf distribution [Kapitsa, 1997: p. 112]. Numerous examples of the Zipf distributions in the inorganic world are given in [Kudrin, 1995].

The distributions of individuals of given species on the Earth surface are non-Gaussian. We have built four distributions using the data taken from the book [Odum, 1986]: (1) the rank distribution of 8 square areas of 0.1 hectares on Earth’s surface according to the number of wolf spiders Lycosa timuqua (the Zipf distribution with \( \alpha \approx 0.7 \)), (2) the rank distribution of 4 square areas of 0.1 hectares on Earth’s surface according to the number of wolf spiders Lycosa carolinensis (the Zipf distribution with \( \alpha \approx 1.0 \)), (3) the rank distribution of 6 square areas of 0.1 hectares on Earth’s surface according to the number of wolf spiders belonging to 3 species, including two previous (the Zipf distribution with \( \alpha \approx 1.01 \)), (4) the rank distribution of 22 species of birds by their prevalence in the pine forest (the Zipfian distribution with \( \alpha \approx 1.0 \)).

The distributions of plant species in the forest by their biomass (the domination-diversity curves) are non-Gaussian. Yu. Odum [Odum, 1986: p. 132] gives such distributions for the four types of forests.

Today, it is getting clear that the prevalence of non-Gaussian distributions is closely related to the fact that the observable world [Haitun, 2005: p. 154–157] is fractal. Non-Gaussian nature of distributions is the immanent property of fractals.

Let us demonstrate how fractals generate non-Gaussian distributions. Here we talk about distributions of fractal (sub)systems by their sizes and distances between them. The fractal dimension is defined by [Mandelbrot, 1977: p. 43]

\[
D \approx \frac{\ln M(\varepsilon)}{\ln(1/\varepsilon)}. \tag{10}
\]

Here \( \hat{I}(\varepsilon) \) is the number of measurement \( D \)-dimensional «cubes», by which we cover a fractal set when determining its measure; \( \varepsilon \) is the edge length of a «cube».

From this expression we obtain the number of substructures \( n(x) \)

\[
n(x) \sim \frac{1}{x^D}, \tag{11}
\]

where \( x \) is a substructure linear size. This is the Zipf distribution with \( \alpha = D - 1 \).

In the expression (10) \( \varepsilon \) is replaced by \( x \) and \( M(\varepsilon) \) is replaced by \( n(x) \). The smaller the parameter \( \alpha \), and, consequently, \( D \), the more this distribution is non-Gaussian. Since, in spite of the point of view coming from the father of fractals B. Mandelbrot, the fractal dimension \( D \) is less than the topological dimension of the fractal (the dimension of the space in which it is placed), the dimension \( D \) of fractals, located in our three-dimensional space, is less than 3. Therefore, the parameter \( \alpha \) of the Zipf distributions generated by these fractals is less than 2, and these distributions are essentially non-Gaussian. (One can read about the Zipf distributions generated by fractals in [Timashev, 1995; Malinetsky, Potapov, 1996]).

It is surprising that in the observed world together with non-Gaussian distributions there are many Gaussian distributions, especially in the natural environment surrounding us. For example, distributions of individuals by weight, size and by many other parameters are Gaussian. It seems that this is largely due to gravity acting on the Earth, which limits sizes of terrestrial objects. In general, the
distribution width, i.e. the degree of their non-Gaussivity, is determined not only by gravity but also by other interactions. Provided these external constraints are absent, distributions of objects are non-Gaussian, as is, for example, the distribution of cosmic bodies by mass, mentioned in this section.

As you know, in the process of evolution entropy is increasing. In the case of a single distribution we see the following: the wider the distribution (hence its form is more simple) the more its entropy. Distributions with greater entropy have longer tails, i.e. they are more non-Gaussian, that corresponds to greater variety of forms. Thus, the evolution towards distributions with increasing entropy means the evolution towards more and more non-Gaussian phenomena. More precisely, we can say that during the evolution a share of distributions with greater entropy is increasing. In other words, with the course of evolution general populations of distributions are getting more non-Gaussian.

It seems in particular, that in the organic world a proportion of mental factors is lower than in the social life. I believe that if for the organic world a random sample of stationary distributions was built and for each such distribution the parameter $\alpha$ of a Zipfian distribution was obtained, than the "organic" distribution of the parameter $\alpha$ would be less non-Gaussian than the corresponding distribution which we have built for the social world with $\alpha \approx 1.31$ (see . Fig. 6).

In broader terms: if for all distributions of the observed world we could build a distribution of values of the parameter $\alpha$ characterizing these distributions, then with the course of evolution this distribution would become more and more non-Gaussian, i.e. characterized by greater value of entropy and lower value of $\alpha$.

Now we should return to the subject of this article. In spite of prevalence of non-Gaussian distributions, I think that all of them describe macro systems. For systems, the scale of which is characteristic for human beings (and, presumably, for hypothetical intelligent beings on other planets), i.e. for mesoscale systems, in my opinion, we can identify quite confidently Gaussian systems with natural world and non-Gaussian systems with artificial objects.

In support I will tell on the results of the statistical analysis [Haitun, 1988] of the article published by one of my critics, M. Kunz [Kunz, 1988]. But first, a few words about the “ideology” of this analysis.

Distributions of people by weight, height, etc., as already mentioned, are Gaussian, while distributions by creative abilities, preferences, etc. are non-Gaussian (that is why there is a phenomenon of non-Gaussian nature of social phenomena). Neural structure of the human seems to generate exclusively non-Gaussian distributions with small values of the parameter $\alpha$ (the order of unity). Gaussian components in human activities are explained, in my opinion, by natural and physiological environment in which a human brain operates. The degree of non-Gaussivity of a distribution is determined by the degree of brain’s involvement in a mixture of factors that determine its (distribution) form.

We have built four distributions analyzing the paper by M. Kunz (Figs. 11–14).
Fig. 11. Distribution of words in Kunz’s article by the number of letters in a word (mathematical symbols were not taken into account). Logarithmic coordinates.

Fig. 12. Distribution of sentences in Kunz’s article by the number of words (mathematical symbols were taken into account). Logarithmic coordinates.
Fig. 13. Distribution of words in Kunz’s article by the frequency of occurrence: (a) frequency form; (b) rank form. Logarithmic coordinates.

Fig. 14. Rank distribution of words with 3 letters by the frequency of occurrence in Kunz’s article. Logarithmic coordinates.
These distributions are presented, as I believe, in the order of increasing the degree of brain’s participation in their formation. A number of letters in the word is limited largely by the vocal apparatus, so that the role of the brain in the formation of the first distribution is relatively small. The number of words in a sentence is determined by the vocal apparatus to a lesser extent. It seems that the role of the brain is more important here. The form of the distribution according to the frequency of occurrence of words is already largely determined by the brain while the role of the vocal apparatus reveals here only by relative frequency of using words of different lengths. This restriction has been removed in the last distribution. Thus, if our hypothesis is correct, these four distributions are constructed in the order of increasing degree of their non-Gaussian nature, i.e. in the order of decreasing $\alpha$.

It is easy to see that the parameter $\alpha$ is decreasing.

9. The second pitfall: animals also generate non-Gaussian distributions in their activity

It is interesting that animal’s behavior is no less non-Gaussian than human’s behavior. Having used the data given in the monograph [Resnikova, 1883], we have constructed distributions characterizing ant’s behavior [Haitun, 1998: p. 145–146]. It turns out that they are the Zipf distributions with the characteristic exponent $\alpha < 1$ (see Figs. 15 and 16).

![Fig. 15. The rank distribution of three ant species by the rate of their research activity determined by the ratio of the time spent on inspection of models (underground labyrinth, brush, parallel plates) of real natural situations (underground passages, crevices between the stones, grass thickness), to the time spent on staying on the exposed surface. The logarithmic coordinates.](image-url)
Fig. 16. The rank distribution of various elements of ant’s behavior (intermittent running, orienteering with the survey of “hills”, detailed survey of small areas of soil, etc.) of a certain kind of ants by the percentage of time for the behavioral ensemble “near-nest exits”. The logarithmic coordinates.

Note that, for example, the distribution of scientists by the number of publications is described by the Zipf distribution with the parameter $\alpha = 1$, and the distribution of scientific slang words by the frequency of their occurrence in §§ 6 and 9 of Einstein’s article on the special theory of relativity (2005) is the Zipf distribution with $\alpha = 0.43$ [Haitun, 1983: p. 258, 275].

It is impressive that the brain of animals, including insects, generates distributions, comparable to the degree of non-Gaussivity with distributions generated by human beings. Presumably, this means that in the social and organic areas of human’s and animal’s activity these non-Gaussian distributions are generated by neural structures.

Thus, if on a particular planet (on the Earth or any other) an object of unknown origin described by a non-Gaussian distribution would be detected, it could mean that this object could be created either by a highly developed civilization or by animals as well. However, this suggestion cannot be applied to the cosmic radio signals because it is difficult to imagine that animals are able to send radio signals into the space.

10. The third pitfall: the rank distortion effect

In contrast to the frequency form of the Zipf distribution (1), there is the factor B in the rank differential form of the Zipf distribution (9). Due to this factor the distribution (9) can deviate from the straight line for the first ranks, i.e. at large values of the random variable $x$. The larger the sample value of B, the greater is the deviation (see Fig. 17)
The rank distortion effect is necessary to take into account while determining empirical values of the parameter $\alpha$ and while determining whether a given distribution is Gaussian or not. As one can see, this effect has purely mathematical nature, and it is calculable with help of the expression for the factor $B$ given by (9).

It can be shown on the well-known Bradford distribution. To approximate this distribution, for the first time the Zipf distribution was used [Haitun, 1983: p. 388–390] to take into account the rank distortion effect (see Fig. 18).
11. Summary

The statistical criterion discussed here (the dichotomy Gaussian / non-Gaussian distributions) can be used for recognition of objects of artificial origin (terrestrial or extraterrestrial), although not with absolute reliability. However, one hundred percent reliable statistical criteria do not exist.

References


