ISSN 0005-1055, Automatic Documentation and Mathematical Linguistics, 2013, Vol. 47, No. 5–6, pp. 169–179. © Allerton Press, Inc., 2013.
Original Russian Text © V.V. Neshitoii, 2013, published in Nauchno-Tekhnicheskaya Informatsiya, Seriya 2, 2013, No. 6, pp. 00–00.

Methods for Calculating the Boundaries of the Core and Zones of Scattering of Publications

V. V. Neshitoii

The Belarus State University of Culture and Arts, Minsk, Belarus

e-mail: neshitoi_vv@tut.by

Received May 3, 2013

Abstract—Three approaches to calculating the boundaries of the core and zones of scattering of publications, and more specifically, the coordinates of three characteristic points, viz., A, C, and B, on the distribution curve are investigated. These are the analytical and graphical methods, as well as the method of least squares. The first two methods can be applied to any statistical rank-size distribution in the case of a homogeneous sample. Such distributions can be described by the second system of continuous distributions, which represents the universal law of scattering of publications. In a particular case, if Weibull's law can be met, the method of least squares is applied. Practical examples that confirm the high accuracy level of statistical rank-size distributions approximated by a second system of continuous distributions are presented.

Keywords: three characteristic points of distribution curves, boundaries of the core and scattering zones, analytical method, graphical method, the method of least squares, a second system of continuous distributions, presentation of rank-size distributions, the universal law of scattering of publications

DOI: 10.3103/S0005105513060034

INTRODUCTION

Statistical rank-size distributions are widely applied in various scientific studies for the description of different objects: periodicals, books, and other publications grouped in a decreasing order of frequency, lexical units of frequency dictionaries, keywords grouped in the process of document indexing in a decreasing order, and many others. The index numbers of journals, books, or words included in such lists are called the ranks. The use of statistical rank-size distributions allows one to solve a wide range of practical problems, including the calculation of the boundaries of the core and scattering zones of publications. Unfortunately, analysis of a variety of scientific papers referring to rank-size distributions reveals the failure of authors to calculate the theoretical law, as well as the boundaries of the core and scattering zones. This can only be achieved by using the theory of generalized distributions.

In 1948, Samuel C. Bradford formulated a law of scattering for journal publications, which says that "If scientific periodicals can be grouped according to a descending number of papers on a particular subject they contain, in the obtained list one can identify the core of journals related to the subject, as well as several groups or zones, each of which contains the same number of papers as the core. In that case, the numbers of journals in the core and the subsequent zones interrelate as follows: $1 : n : n^2$" [1, p. 93, 94].

However, this thesis does not shed light on how to calculate the core and zones of scattering by using statistical rank-size distributions because it does not provide any formula for calculations. It is also not clear which theoretical rank distribution is taken for the basis of the Law of Scattering, which points on the rank-size distribution plot are taken for the boundaries, how many zones of scattering can exist, and how to calculate the value of n. Bradford's statement only refers to the fact that it is possible to establish a core of journals and several zones. It is also assumed that the number of articles in each zone is equal to the number of papers in the core. However, this assumption is inconsistent with factual evidence and is not grounded theoretically.

METHOD OF SELECTION

Taking Bradford's thesis as a basis, researchers have to define the volume of the core of journals as they deem fit, whereas the number of zones of scattering and their size are obtained on the basis of factual evidence derived from the condition that each zone contains the same number of papers as the core. Under this approach, the number of zones of scattering for the same statistical rank-size distribution can vary greatly among different researchers.

The main idea of Bradford's Law that cannot be doubted is related to the grouping of journals according to a diminishing number of papers published on a

particular subject, i.e., ranking. Rank-size distributions most comprehensively reflect the nature of scattering of publications. However, many researchers avoid addressing this key issue. As a result, they try to establish the core and zones of scattering based on Bradford's thesis of 1948 instead of trying to study the statistical properties of rank-size distributions and a theoretical distribution law with the same properties.

Unfortunately, the literature sources do not refer to any statistical rank-size distributions of book titles grouped in a descending order of frequency of occurrence. Furthermore, the work on the compilation of frequency lists of journals organized on the basis of various criteria is not actively pursued. Yet, mathematical linguistics can be applied to study the properties of rank-size distributions. This field of knowledge has accumulated a large number of frequency dictionaries in which the lexical items are grouped in a descending order (or rather, a non-increasing order) of frequency of their text usage. A word-frequency table was created in order to facilitate the analysis. Such a table contains the necessary information about the frequency dictionary, such as the rank of a word or a range of ranks, the absolute frequency of word usage, the number of words with a given frequency, cumulative absolute frequency, relative frequency, and the cumulative relative frequency of word usage. A similar table is usually annexed to a frequency dictionary. It allows one to present the frequency structure of a dictionary of any size in a rather concise way.

The approximation of statistical distributions, including rank-size ones, involved the creation of universal mathematical models, viz., generalized distributions, as well as the development of methods for calculating the distribution laws based on statistical data and the related estimated parameters. A new form of presentation of rank-size distributions, which laid the basis for homogeneity criteria and the adequacy of a sample size, was introduced. Special software, including free computer programs, can also be used for the construction of frequency dictionaries.

This means that many *tasks related to processing statistical rank-size distributions in linguistics, computer science, and library science were already solved long ago*. Unfortunately, the accumulated knowledge in these areas and the results of related scientific research are practically not used. The reason is that researchers have to make significant time-consuming efforts in order to study and apply these results in their practical scientific work. It is much easier to formulate traditional hypotheses and test them according to various criteria of fitness. However, one cannot create new knowledge by following this path.

The above-mentioned method for the selection of the size of the core and zones of scattering is similar to the method of hypotheses formulation, which, however, does not involve checking the results by testing a fit. Such a method is not acceptable in scientific research.

Analytical Approach

The rank-size distribution of journals is given. If one can find a theoretical distribution that accurately describes the statistical rank-size distribution, it will be possible to formulate with mathematical precision the Law of Scattering in Bradford's interpretation. In other words, this law should be derived from the rank-size distribution as a particular case. However, *only the theoretical rank-size distribution can serve as the most common universal law of scattering* given that the distribution law is the most comprehensive characteristic of any random variable.

In order to define the boundaries of the core zone and zones of scattering, many authors try to process statistical rank-size distributions without solving the general problem related to the establishment of the law of distribution. For this purpose, in order to approximate each of the statistical distributions researchers have to try to apply different formulas instead of using one theoretical law with different variables. In addition, the selection of a theoretical curve is performed by using an inconvenient form of statistical rank-size distributions in the "log rank-log frequency" system of coordinates, which contains insufficient data about the rank-size distribution and, moreover, has no probabilistic meaning.

Therefore, the problem of scattering of publications can be solved on the basis of the solution to a more general problem, viz., a universal formula or a system of formulas that can accurately approximate a variety of rank-size statistical distributions. Given that " ... scattering of scientific data is the cornerstone of all scientific information activities, and the study of this property of scientific information is the major challenge of computer science..." [1, p. 93], this problem needs to be solved by very serious means such as the generalized distributions described by the author in [2–6].

Let's consider the first and the second systems of continuous distributions, each of which is defined by three generalized densities. The first density of these systems can be presented as:

$$p(x) = Ne^{k\beta x}(1 - \alpha u e^{\beta x})^{\frac{1}{u}-1}, \quad (1)$$

$$p(t) = Nt^{k\beta-1}(1 - \alpha u t^{\beta})^{\frac{1}{u}-1}. \quad (2)$$

In this case, α, β, k, and u are the distribution parameters calculated by using the statistical distribution. N is a normalizing factor, which is expressed in the distribution parameters, provided that the area below the distribution curve is equal to 1.

The first density has one remarkable property according to which under $u \le 0.5$ parameters values the first density *has the x_C mode*, that is, the value of a random variable X, wherein the $p(x)$ density is maximum, and two *inflection points*, viz., x_A and x_B – are equally distanced from both sides of the mode. The points on the density-distribution plot separate the convex portion from the concave curve and the concave part from the convex portion. When $1/2 < u < 1$, there are only two specific points, viz., x_A and x_B (for distributions with left-sided asymmetry). When $u \ge 1$, there are no characteristic points. Distribution (1) can be defined on the whole real axis, i.e. $-\infty < x < \infty$.

Distribution (2) is set on the positive half axle $t > 0$. The $p(t)$ density plot can have various forms. For certain values of the parameters k, β, and u the density plot $p(t)$ can have the form of a decreasing distribution curve, which means that the density $p(t)$ describes not only the single-vertex statistical distributions, but also rank-size (decreasing) distributions.

Two methods were elaborated in order to calculate the distribution law of random variables by using statistical distributions, including the rank-size one: the general method of moments and stable general method [4–6]. In order to simplify the calculations and to extract new data from the rank-size distribution, the latter is transformed into a single-vertex distribution [7]. This is achieved by reducing the form of the second density to the form of the first density. For this purpose, the left and right sides of the $p(t)$ density are multiplied by t, and t^β is represented in the form of $e^{\beta \ln t}$. Density (2) will look like as follows:

$$tp(t) = Ne^{k\beta \ln t}(1 - \alpha u e^{\beta \ln t})^{\frac{1}{u} - 1}. \quad (3)$$

Comparing (1) and (3), the following can be established:

$$tp(t) = p(x), \quad \ln t = x, \quad (4)$$

Thus, considering the equations (4), the rank-size distribution (2) converted to the form of $tp(t) = f(\ln t)$ represents the distribution (1) and possesses all its properties. Specifically, it has the $\ln t_C$ mode and two inflection points, $\ln t_A$ and $\ln t_B$. This information about the rank-size distribution is new. *Let's take the abscissae of the points A, C, B of generalized distributions for the boundaries of the core and zones of scattering* of various objects.

As the $\ln t_A$ and $\ln t_B$ points are at equally distances from the $\ln t_C$ mode, the equation can be presented as follows: $\ln t_C - \ln t_A = \ln t_B - \ln t_C$, so that $\ln(t_C/t_A) = \ln(t_B/t_C)$, or

$$\frac{t_C}{t_A} = \frac{t_B}{t_C} = n. \quad (5)$$

Equation (5) can serve as a refined interpretation of the law of scattering of publications in Bradford's tradition, although it provides a new interpretation of the law of scattering of publications.

Based on (5), two other formulas can be introduced:

$$t_A : t_C : t_B = t_A(1 : n : n^2), \quad (6)$$

$$t_A : t_I : t_{II} = t_A[1 : (n-1) : n(n-1)]. \quad (7)$$

In this case, t_A, t_C, t_B are the number of journals ranging from the beginning of the frequency list to the A, C, and B points; t_I and t_{II} are the number of journals in the first and the second zones of scattering. Based on these equations, formula (7) is obtained from (6).

All three interpretations described using formulas (5)–(7) differ from Bradford's interpretation and further clarify Bradford's thesis. However, the *above formulas* (similarly to Bradford's law) do not represent the law of scattering as they do not provide all the necessary information. These formulas only establish a general relationship between the abscissae of the three characteristic points, viz., the mode and the two inflection points. Yet, the calculation of the coordinates of these points requires the knowledge of the rank-size distribution law. The information required for the calculation of the boundaries of the core of journals and zones of scattering, the share of articles in the core and zones of scattering, as well as the share of articles published in the journals on top of the frequency list, can only be obtained if specific parameters of this distribution are known. This data is expressed in the distribution function. However, formulas (5)–(7) do not apply the distribution function.

From this point of view, the rank-size distribution law provides the most comprehensive information. This means that the generalized density (2), or rather, the *second system of the continuous distributions defined by three generalized densities, represents the universal law of scattering of publications*. It is appropriate to note that the first system of continuous distributions, including density (1), is a universal law of publication aging [2].

A number of researchers argue that Bradford's law is Zipf's law, viz., $p_r = k/r$ presented differently. This formula was proposed by Zipf in order to describe the rank-size distribution of words in a frequency dictionary. Thus, p_r is the relative frequency of a word with the r rank, and k is a parameter. For Zipf's law, multiplication of the relative frequency of a word by its rank is equal to the k constant, which is depicted as a horizontal line without any specific points on the plot in the $rp_r = f(\ln r)$ system of coordinates. This demonstrates that Zipf's law cannot approximate statistical rank-size distributions because the latter have the form of a unimodal curve with two inflection points in case of a homogeneous sample in the given system of coordinates. Such curves can be accurately described by the second system of continuous distributions. It should be noted that Zipf's law, and in a more general

interpretation with two additional parameters the law of Estu–Zipf–Mandelbrot is a particular case of the generalized distribution (2) if $u = 1$, $\beta < 0$. However, under this condition, the distribution curve $rp_r = f(\ln r)$ has no characteristic points. Therefore, it is neither possible to develop the scattering law on their basis, nor to approximate statistical rank-size distributions, especially in serious scientific studies. The best theoretical distribution for the approximation of the statistical rank-size distribution needs to be calculated based on a second system of continuous distributions [6].

We next consider Bradford's law of scattering (where $r = t$, $t_N = t_A$)

$$t_N : t_I : t_{II} = t_N(1 : n : n^2). \quad (8)$$

Comparing this formula to (6) and (7), one can see that Bradford's law of scattering is an incorrect combination of two correct formulas: the left side corresponds to the formula (7), and the right to the formula (6).

The equation (8) can be resented differently, taking the three characteristic points into account, which were not mentioned by Bradford in his version of the law of scattering. From (8) we have

$$t_N : (t_C - t_N) : (t_B - t_C) = t_N(1 : n : n^2).$$

t_N is moved out of the brackets in the left part of the equation

$$t_N[1 : (t_C - t_N)/t_N : (t_B - t_C)/t_N] = t_N(1 : n : n^2).$$

Next, we can write two equations $(t_C - t_N)/t_N = n$ and $(t_B - t_C)/t_N = n^2$, from which we can establish that the relationship $t_C : t_N = n + 1$, and the relationship $t_B : t_N = n^2 + n + 1 = (n + 1)^2 - n$. As a result, the following formula is obtained:

$$t_N : t_C : t_B = t_N[1 : (n + 1) : (n^2 + n + 1)]. \quad (9)$$

In our precise formula (6), the first relationship is equal to n, and the second is equal to n^2. Since these two relationships are not met in (9), according to Bradford's law, the inflection points of the rank-size distribution curve $tp(t) = f(\ln t)$ should be located at *different distances* from the $\ln t_C$ mode, which is contradictory to the statistical and theoretical properties of the rank-size distributions. Another contradiction to Bradford's law consists in the assertion that each zone of scattering contains the same number of papers as the core. Naturally, these two requirements cannot be satisfied by any theoretical distribution. Therefore, all attempts to improve Bradford's law of scattering while maintaining its inherent contradictions have been unsuccessful. This is not surprising, because the attempts were made to solve a particular problem without solving the general problem related to the establishment of the rank of such a theoretical distribution that would allow one to accurately approximate a wide variety of statistical rank-size distributions.

Let's estimate the accuracy of formula(9). Let's take $n = 5$ and calculate $t_C : t_N$ and $t_B : t_N$. In the first case, the relationship is equal to $n + 1 = 6$ and in the second $n^2 + n + 1 = 31$. Based on the accurate formula (6), 5 and 25 are obtained, respectively. Error calculation of the abscissae of the points C and B was 20% and 24%, respectively. The size of the core is taken as equal in both cases, because it is impossible to calculate it based on Bradford's law. It should be noted that with the increase of n, this error decreases.

The obtained results suggest that Bradford's law of scattering presented in formulas (8) and (9) is relatively close to the exact formulas (6) and (7), although it was formulated without using any theoretical rank-size distribution. Bradford's argument about the same number of articles in the core and zones of scattering is not correct and can be misleading for some researchers.

However, Bradford's great merit lies in the fact that he was the first to draw attention to the phenomenon of the scattering of publications and prompted many researchers to study this phenomenon in-depth.

Yet, this issue has proven to be very challenging. The universal law of scattering was found only after the development of the theory of generalized distributions. It was noted above that such a law is the second system of continuous distributions. Formulas for calculating the boundaries of the core and zones of scattering are derived from the generalized density (2).

The t_C mode is derived from the $dtp(t)/d\ln t = 0$ condition and in the general case for the distributions of types I–V is equal to [4]

$$t_C = \left(\frac{k}{\alpha(1 + ku - u)}\right)^{1/\beta}. \quad (10)$$

The value of n is set by

$$n = \left[1 + \frac{1 - u + \sqrt{[4k(1 + ku - u) + (1 - u)](1 - u)}}{2k(1 + ku - u)}\right]^{1/\beta}. \quad (11)$$

The abscissae of the inflection points are calculated using the following formulas:

$$t_A = t_C/n, \quad t_B = t_C n. \quad (12)$$

The share of articles in each zone and for any other ranking range is calculated by using the distribution function or the statistical rank-size distribution.

As follows from (11) and (12) formulas, for set values of the parameters of the form k and u and with a decrease in the value of the parameter β the value of $n = t_C/t_A = t_B/t_C$ increases. This means that the distribution curve $tp(t) = f(\ln t)$ becomes broader and flatter.

Methods for calculating the approximating four-parameter distributions are presented in several works [2–6]; however, these methods are designed for experienced researchers.

Graphical Method

Boundaries of the core and zones of scattering can be established for the statistical rank-size distribution without preliminary calculation of the theoretical law by using a simple graphical method based on the analysis of the properties of generalized distributions (1) and (2). The essence of this method can be illustrated with the following example.

Let's analyse Weibull's law, in which the distribution function and density are set by the following formulas:

$$F(t) = (1 - e^{-\alpha t^{\beta}}), \quad (13)$$

$$p(t) = \alpha\beta t^{\beta-1} e^{-\alpha t^{\beta}}. \quad (14)$$

If the parameter is $\beta \leq 1$, density (14) can accurately describe some statistical rank-size distributions.

In order to extract useful data of the rank-size distribution, let's transform the density (14) into $tp(t) = f(\ln t)$ [7]:

$$tp(t) = \alpha\beta t^{\beta} e^{-\alpha t^{\beta}} = \alpha\beta e^{\beta \ln t} e^{-\alpha e^{\beta \ln t}}. \quad (15)$$

Given the equality of $tp(t) = p(x)$ and $\ln t = x$, the density (15) will have the following form:

$$p(x) = \alpha\beta e^{\beta x} e^{-\alpha e^{\beta x}}. \quad (16)$$

The resulting formula is a special case of the density (1) where $u \to 0$, $k = 1$. Density (16) provides new information about the rank-size distribution, the density plot, where the distribution curve has three characteristic points, namely the mode and two inflection points, A and B, at equal distances from both sides of the mode. The abscissae of these points are defined. Density (16) is differentiated with respect to x; the first derivative is equated to zero. From this equation the equation for the mode can be established:

$$x_C = \frac{1}{\beta} \ln \frac{1}{\alpha}. \quad (17)$$

At the inflection points the second derivative is zero. Based on this density condition (16), the following can be established:

$$x_A = x_C - \frac{1}{\beta} \ln \frac{3 + \sqrt{5}}{2}, \quad x_B = x_C + \frac{1}{\beta} \ln \frac{3 + \sqrt{5}}{2}.$$

Let's introduce the notation

$$n = \left(\frac{3 + \sqrt{5}}{2}\right)^{\frac{1}{\beta}}. \quad (18)$$

Taking (18) into account, the last two formulas can be presented in a simpler form:

$$x_A = x_C - \ln n, \quad (19)$$

$$x_B = x_C - \ln n. \quad (20)$$

Based on the Weibull distribution, using formulas (19) and (20) and the equation $x = \ln t$ one can establish that

$$\ln t_C - \ln t_A = \ln t_B - \ln t_C = \ln n,$$

where $\ln(t_C/t_A) = \ln n$, or

$$\frac{t_C}{t_A} = \frac{t_B}{t_C} = n. \quad (21)$$

Here

$$t_C = \left(\frac{1}{\alpha}\right)^{\frac{1}{\beta}}, \quad t_A = \frac{t_C}{n}, \quad t_B = t_C n. \quad (22)$$

Formula (21) established on the basis of Weibull's law coincides with the analogous formula (5), which was obtained by using the four-parameter distribution (2). This formula is also valid for any particular case of distribution (2), when the parameter of the form is $u \leq 1/2$. Weibull's law is a special case of the Weibull distribution (2) for $u \to 0$, $k = 1$. In some cases it accurately describes rank-size statistical distributions. However, the generalized density (2), which allows one to calculate both the coordinates of characteristic points and the distribution function [4, pp. 155–160] for virtually any statistical rank-size distribution, remains as a universal law. In some cases, statistical rank-size distributions can be accurately described with additional densities of the second system of continuous distributions.

Let's assume that the rank-size distribution is set by Weibull's law with the parameters of $\alpha = 0.1$; $\beta = 0.5$. It is transformed into the density $p(x)$ presented by the formula (16). In this case, the formulas (17)–(20) allow one to establish the following: $n = 6.8541$; $x_C = 4.6052$; $x_A = 2.6804$; and $x_B = 6.53$.

The values of the $p(x)$ density with an interval of $\Delta x = 0.5$ are calculated based on formula (16) and the results are summarized in Table 1.

The density $p(x)$ plot, i.e., the distribution curve, is built (Fig. 1a).

On the distribution curve, the abscissa of the C point can easily be defined graphically with a horizontal tangent to the curve (Fig. 1a).

In order to find the abscissae of the inflection points the properties of the distribution curve are applied, according to which, the first derivative has its extreme values at the points A and B: it has the maximum value at the A point and the minimum value at the B point. The slope of the curve segments to the horizontal axis at the interval borders is calculated as the ratio of the difference between the values of the $p(x)$ density at the boundaries of the interval to the interval width Δx. In other words, the approximate values of the first derivative at the midpoints of the intervals are calculated (Table 1 shows the calculated values of the derivative $dp(x)/dx$) and the plot is constructed (Fig. 1b).

Table 1. The values of the density and the slope of the tangent to the curve at the midpoints of the intervals

x	$p(x)$	$dp(x)/dx$	x	$p(x)$	$dp(x)/dx$
−2	0.01773	0.008539	4	0.176464	0.023037
−1.5	0.022529	0.010732	4.5	0.18369	0.004705
−1	0.028542	0.013405	5	0.180147	−0.01966
−0.5	0.036022	0.016609	5.5	0.163655	−0.04617
0	0.045242	0.020359	6	0.134756	−0.06795
0.5	0.056465	0.024607	6.5	0.097806	−0.07722
1	0.069906	0.02919	7	0.060369	−0.06977
1.5	0.085655	0.033761	7.5	0.030263	−0.04921
2	0.103565	0.037706	8	0.011614	−0.0259
2.5	0.123099	0.040067	8.5	0.003163	−0.00951
3	0.143144	0.039496	9	0.000554	−0.00222
3.5	0.161833	0.034352	9.5	5.52E−05	−0.00029

The first derivative $dp(x)/dx$ has its maximum at the x_A point and its minimum in the x_B point. These points can be easily defined with the horizontal tangent to the curve in Fig. 1b.

The graph allows one to establish the approximate value of the abscissae of the three characteristic points: $x_A = 2.7$; $x_C = 4.6$; and $x_B = 6.5$.

Regarding the rank-size distribution, the following can be established: $t_A = \exp(x_A) \approx 15$; $t_C = \exp(x_C) \approx 99$; $t_B = \exp(x_B) \approx 665$. The exact values of the Weibull distribution are $t_A = 14.59$; $t_C = 100$; $t_B = 685$.

This simple method enables one to define the abscissae of three characteristic points of any statistical rank-size distribution without calculating the theoretical distribution law, but with the use of its properties.

The values of the distribution function $F(t)$ for any given value of the rank t, including in three character-

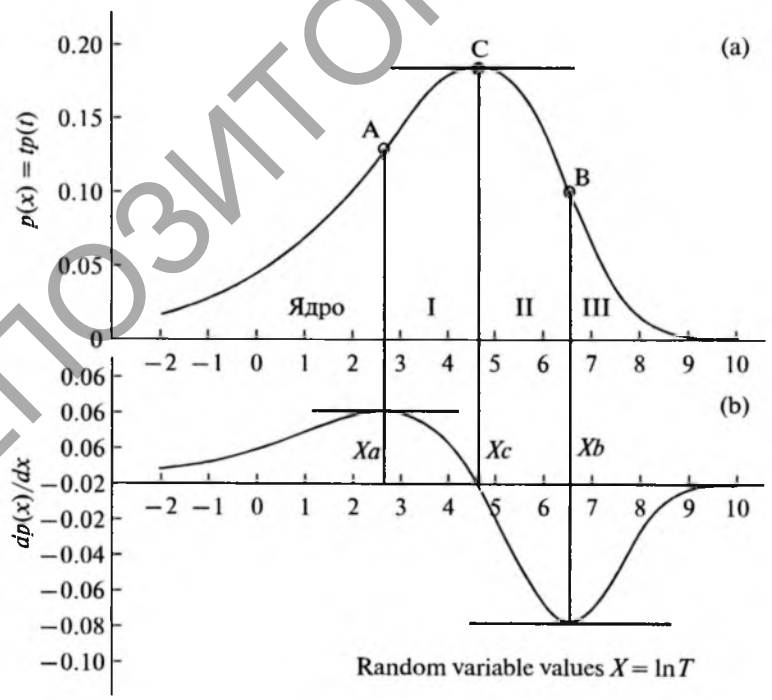


Fig. 1. Distribution density (a) and first derivative distribution (b) graphs.

istic points, can be calculated by using the statistical rank-size distribution.

Using the graphical method, different researchers will have similar results in terms of the defined boundaries of the core and zones of scattering for the same rank-size statistical distribution.

It should be noted that, as presented in Fig. 1a, the theoretical distribution curve smoothly reaches its maximum value and then gradually decreases. Therefore, the calculation of the slope of the curve segments on the intervals is rather simple. A similar curve of the statistical distribution $tp(t) = f(\ln t)$ has many peaks and troughs, which makes the construction of Fig. 1b rather difficult. Therefore, it is necessary to preliminary smoothen it, for example, by using a drawing curve.

It is also necessary to note that in Fig. 1 the horizontal tangent to the curve of the distribution of $tp(t) = f(\ln t)$ represents Tsipf's law at point C. It follows that this law cannot be described by any rank-size distribution

Least-Squares Method

In some cases, the statistical-rank distribution can be accurately described by Weibull's law, whose distribution function and the probability density are set by (13) and (14). This law was first used by G.G. Belonogov in order to describe the rank-size distribution of words in a frequency dictionary [8]. This particular law is very simple; therefore, it is advisable to test it in the first place while searching in a suitable rank-size distribution. For such a test, the distribution function must be converted to a linear form

$$\ln\ln(1/(1-F(t))) = \ln\alpha + \beta\ln t. \quad (23)$$

The following notations are introduced:

$$Y = \ln\ln(1/(1-F(t))), \quad X = \ln t. \quad (24)$$

In that case, the last equation can be written as

$$Y = \ln\alpha + \beta X. \quad (23')$$

In order to test the applicability of Weibull's law, it is necessary to calculate the values of X and Y by using the formulas (24) and to create a dependence plot $Y = f(x)$ based on the statistical distribution function. If the empirical points are scattered along the line (23′), one can calculate the values of α and β parameters of the line based on the method of least squares:

$$\beta = \frac{\overline{XY} - \bar{X}\bar{Y}}{\overline{X^2} - (\bar{X})^2}, \quad \alpha = \exp(\bar{Y} - \beta\bar{X}). \quad (25)$$

To evaluate the closeness of the linear relationship between the Y and X variables, the sample correlation coefficient is calculated

$$R_{y/x} = \frac{\overline{XY} - \bar{X}\bar{Y}}{\sigma_x\sigma_y}, \quad (26)$$

where the average standard deviations σ_x and σ_y are:

$$\sigma_x = \sqrt{\overline{X^2} - (\bar{X})^2}, \quad \sigma_y = \sqrt{\overline{Y^2} - (\bar{Y})^2}.$$

Where

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i, \quad \bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i, \quad \overline{X^2} = \frac{1}{N}\sum_{i=1}^{N} X_i^2,$$

$$\overline{Y^2} = \frac{1}{N}\sum_{i=1}^{N} Y_i^2,$$

where N is the number of random variables X, Y.

The abscissae of points A, C, and B for Weibull's law are calculated according to (18) and (22). The values of the distribution function at these points for any values of α and β parameters are equal to, respectively:

$$F(t_A) = 0.31748, \quad F(t_C) = 0.63212, \quad F(t_B) = 0.92705. \quad (27)$$

One should note that the statistical data is scattered along a straight line in the case of a homogeneous sample for which Weibull's law holds. However, any attempt to describe the rank-size distribution of words in a frequency dictionary will show that the first 50–100 most frequent words do not follow Weibull's distribution. They are basically part of the heterogeneous part of the sample. Therefore, in order to achieve a more accurate description of the rank-size distributions one can preliminary delete the first 50–100 words with the subsequent recalculation of ranks and the relative frequencies of words, which will result in a more homogeneous sample. If it is necessary to approximate the rank-size distribution of all words in a frequency dictionary, including the functional words, one should introduce an additional parameter to the theoretical distribution. Our previous research showed that Weibull's law can be expressed as follows in view of the third parameter (denoted by δ) [9]

$$F(t) = 1 - e^{-\alpha[(t+1)^{\beta} - e^{-\delta t}]}, \quad (28)$$

$$p(t) = \frac{\alpha\beta(t+1)^{\beta-1} + \alpha\delta e^{-\delta t}}{e^{\alpha[(t+1)^{\beta} - e^{-\delta t}]}}. \quad (29)$$

The α and β parameters can be calculated using the method of least squares according to formulas(23)–(25) for the ranks of frequency dictionary words ranging from 50–100 to the ranks of words with frequency of 2–3. Furthermore, an additional parameter δ is calculated on the basis of estimated parameters by using the formula which is obtained from the distribution function (28):

$$\delta = -\frac{1}{t}\left[\ln\left((t+1)^{\beta} - \frac{1}{\alpha}\ln\frac{1}{1-F(t)}\right)\right]. \quad (30)$$

Table 2. Calculation of the Weibull-law parameters according to statistical distribution

Word ranks	Distribution function	$\ln r$	$\ln\ln\frac{1}{1-F(r)}$					
r	$F(r)$	X	Y	XY	X^2	Y^2	$Y@$	$F@$
80	0.354324	4.382027	−0.82678	−3.62295	19.20216	0.683558	−0.83551	0.351864
100	0.374545	4.60517	−0.75656	−3.48411	21.20759	0.57239	−0.76646	0.371648
150	0.411675	5.010635	−0.63398	−3.17665	25.10647	0.401932	−0.641	0.409488
250	0.458992	5.521461	−0.48724	−2.69026	30.48653	0.2374	−0.48294	0.460423
400	0.506106	5.991465	−0.34894	−2.09067	35.89765	0.12176	−0.3375	0.510098
666	0.561671	6.50129	−0.19263	−1.25236	42.26677	0.037107	−0.17975	0.566333
1097	0.620558	7.000334	−0.03144	−0.22006	49.00468	0.000988	−0.02533	0.622802
1809	0.681615	7.500529	0.134963	1.012291	56.25794	0.018215	0.129442	0.679603
3000	0.740589	8.006368	0.299617	2.398842	64.10192	0.08977	0.285962	0.735798
4900	0.792717	8.49699	0.453411	3.852626	72.19885	0.205581	0.437774	0.787594
8100	0.838483	8.999619	0.600563	5.404838	80.99315	0.360676	0.593301	0.836338
13360	0.875102	9.50002	0.732492	6.958688	90.25039	0.536544	0.748139	0.879133
20000	0.898718	9.903488	0.828485	8.204889	98.07907	0.686387	0.872983	0.90874
268106	0.995819	12.49914	1.700595	21.25597	156.2284	2.892023	1.676148	0.995228
360755	0.997191	12.79595	1.770694	22.65771	163.7364	3.135356	1.767992	0.997146
Sum		116.7145	3.243251	55.2088	1005.018	9.979688		
Mean		7.780966	0.216217	3.680587	67.0012	0.665313		
Beta =	0.309427	Sx =	2.541215					
Alfa =	0.111757	Sy =	0.786488					
Ry/x =	0.999789	δ =	0.091					

The additional parameter can be calculated once for a given relative frequency of the most frequent word, which is equal to the distribution function $F(t = 1)$.

The two-parameter Weibull's law can effectively describe some rank-size distributions of journals, terms, and keywords that form a statistically homogeneous sample, while the three-parameter law is more suitable for the calculations of some non-uniform samples.

Let's analyse the example of the *rank distribution* of words contained in the *Frequency Dictionary of the Modern Russian language* [10]. This dictionary contains a large-scale sample of 135 million expressions. The number of different tokens (referred to as lemmas in the original source) is 739930 items. Of these, the numbers of lemmas with a frequency of two or more, as well as three or more, are 360755 and 268106, respectively. The above source contains the frequency list of top 20000 lemmas, which allows one to calculate the cumulative relative frequencies, i.e., the distribution function for all ranks from 1 to 20000. Two values of the distribution function with a certain number of lemmas with frequencies of usage of one and two times (379175 = 739930 − 92649 = 360755 and 360755 − 268106, respectively) are calculated:

$$F(360755) = 1 - 379175/135000000 = 0.9971913;$$

$$F(268106) = 1 - (379175 + 92649 \times 2)/135000000 = 0.9958187.$$

In this case, the share of lemma usages with the frequency of 1 is subtracted once from the total use of all lemmas equal to one in the first case and two times in the second case.

Table 2 is compiled on the basis of the Frequency Dictionary in order to present the individual ranks of words ($R \geq 80$) as well as the ranks corresponding to the values of the distribution function (see the first two columns). The parameters of Weibull's law and the correlation coefficient are calculated using the least squares method. The β parameter is equal to 0.309427 and the α parameter = 0.111757. The correlation coefficient is $Ry/x = 0.999789$. This means that the empirical relationship is close to the theoretical line (23), as depicted in Fig. 2.

When the estimates of the α and β parameters are known and the distribution function is $F(t = 1) = 0.035802$, it is possible to calculate the third parameter based on the formula (30). The third parameter is required for a more accurate description of the most frequent words: $\delta = 0.091$. Next, the values of the dis-

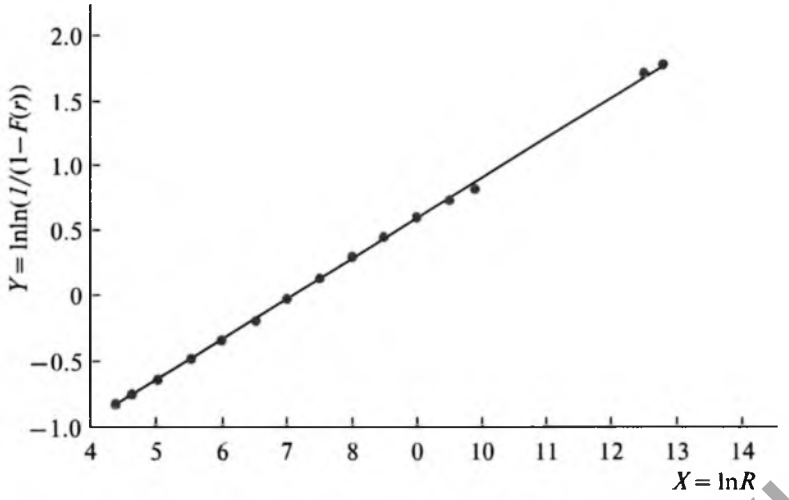


Fig. 2. Weibull's line.

tribution function are calculated according to the three-parameter Weibull's law (Table 3) and a distribution function plot is built at a semi-log scale (Fig. 3), taking the functional words into account, where the empirical distribution function is shown as individual points and the theoretical distribution function is shown as a solid line.

As illustrated by Table 3 and the distribution function plot, the introduction of the third parameter into Weibull's law allows one to approximate the statistical distribution of the first 50–80 words in the Frequency Vocabulary with high accuracy. The rank-size distribution of the remaining words can be described by the classical Weibull's law with two parameters.

The abscissae of the three characteristic points are calculated according to the formulas (18) and (22) where the estimated values of the α and β parameters are set: $n = 224287$; $t_C = 1191$, $t_A = 53$, $t_B = 26704$. One should note that the logarithms of the ranks of words in the points A, C, and B are equal to 3.97187, 7.08221, and 10.19256. The theoretical distribution function at these points is equal to 0.31748, 0.63212, and 0.92705.

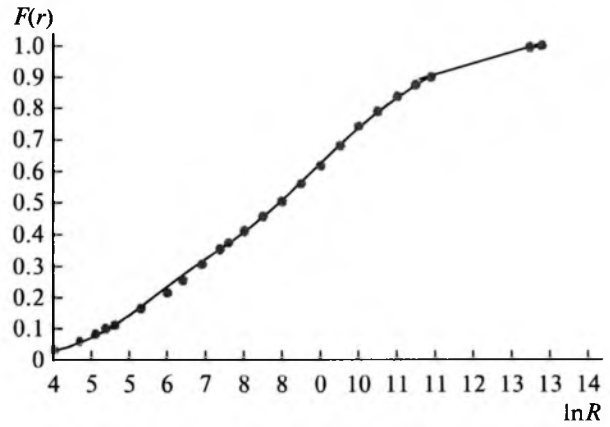
One can see from these calculations that the core of the frequency dictionary is formed by the first 53 words, which cover 31.7% of the text. The first zone of scattering, A – C, contains 1191 – 53 = 1138 words, which cover 31.5% of the text (63.2 – 31.7 = 31.5). The second zone, C – B, contains 26704 – 1191 = 25513 words, which cover 29.5% of the text (92.7 – 63.2 = 29.5). Finally, the third zone of scattering contains the rest of the vocabulary 739930 – 25513 = 714417 words. This huge dictionary only covers 7.3% of the text (100 – 92.7 = 7.3).

Another example can be studied: the *rank-size distribution of periodicals in chemistry and chemical technology* [1]. Because the sample is homogeneous in this case, Weibull's distribution with two parameters can be applied for the approximation of the statistical rank-size distribution. The necessary calculations are performed on the basis of the method described above. The results are presented in Table 4 and Fig. 4.

The obtained results demonstrate Weibull's law allows one to approximate with high accuracy the statistical rank-size distributions of journals grouped in a descending order by the number of published chemistry and chemical technology papers. The correlation coefficient is $Ry/x = 0.999705$. The core contains 88 journals. The number of periodicals up to the C point that are part of the core area and the first zone of scattering is 552. The core and the first two zones of scattering, i.e., up to point B, contain 3469 journals with 92.705% of the articles (out of the total of 187911 papers). The third zone of scattering includes all the

Table 3. Empirical and theoretical distribution function of the most frequently used words

Word ranks	Empirical distribution function	Detailed theoretical distribution function
1	0.035802	0.035798
2	0.067176	0.061847
3	0.085204	0.082922
4	0.101071	0.100782
5	0.113756	0.116317
10	0.165571	0.172804
20	0.217115	0.235534
30	0.255761	0.271023
50	0.309294	0.313454



Dependence function plot: the solid line, calculation

Fig. 3. Weibull's distribution function.

remaining periodicals, 10850 − 3469 = 7381, which contain 100 − 92.705 = 7.295% of the papers.

Despite the high accuracy of approximation of some rank-size distributions by Weibull's law, its application for computer science and mathematical linguistics research is still quite rare. Rather, researchers more often apply Zipf's law, which is not suitable for such studies. Taking into account the fact that both of these laws and many others only constitute special cases of the generalized distribution (2), it is more appropriate to use the second system of continuous distributions for the description of different types of rank-size distribution.

CONCLUSIONS

The calculation of the theoretical distribution law is a major task in processing data series. This task can be rather easily solved by using the methods described in the theory of generalized distributions. The second system of continuous distributions is used to approximate the statistical rank-size distributions.

Based on the analysis of properties demonstrated by generalized distributions, we offer a mathematically precise formulation of Bradford's law of publication scattering. However, such formulations of Bradford's law cannot be accepted as a full-fledged law of publication scattering. The universal law of scattering is provided by the second system of continuous distributions as a generalized four-parametric density, i.e., the law of distribution, most fully characterizes random variables.

A general stable method is used to calculate the law of distribution and estimated values of its parameters for the statistical rank distributions. The abscissae of three characteristic points, viz., A, C, and B, which are accepted by the author as boundaries of the core and zone of scattering zones are calculated according to certain estimated parameters by using preliminary derived formulas. The abscissae of the points C and B that are calculated using Bradford's law and the universal law differ by 20–25% with the proviso that $n = 5$ and the size of the core is the same in both cases. With n increasing, this error decreases.

The use of the analytical method requires at least some basic knowledge about the theory of generalized distributions. This method is designed for trained researchers.

Based on the properties of the rank-size distributions, we have proposed using the graphical method

Table 4. Scattering of journal publications in the field of chemistry and chemical technology (10850 journals, 187911 papers)

The number of journals	The share of articles	$\ln t$	$\ln\ln\frac{1}{1-F(r)}$					
t	$F(t)$	X	Y	XY	X^2	Y^2	$F(t)@$	$Y@$
18	0.15	2.890372	−1.817	−5.25181	8.3542489	3.301489	0.1536	−1.7912
50	0.25	3.912023	−1.2459	−4.87399	15.303924	1.552267	0.24772	−1.25649
100	0.34	4.60517	−0.8782	−4.04426	21.207592	0.771235	0.33577	−0.89371
500	0.62	6.214608	−0.033	−0.20508	38.621354	0.001089	0.61323	−0.05138
1000	0.75	6.907755	0.3266	2.25607	47.717083	0.106668	0.7447	0.3114
2000	0.85	7.600902	0.6403	4.86686	57.773718	0.409984	0.85948	0.67418
Sum		32.13083	−3.0072	−7.25221	188.97792	6.142732		
Mean		5.355138	−0.5012	−1.2087	31.49632	1.023789		
Beta =	0.523374	Sx =	1.67893183	tc =	551.5708	$F(tc)$ =	0.63212	
Alfa =	0.036738	Sy =	0.87896938	ta =	87.69708	$F(ta)$ =	0.31748	
Ry/x =	0.999705	n =	6.28949982	tb =	3469.104	$F(tb)$ =	0.92705	

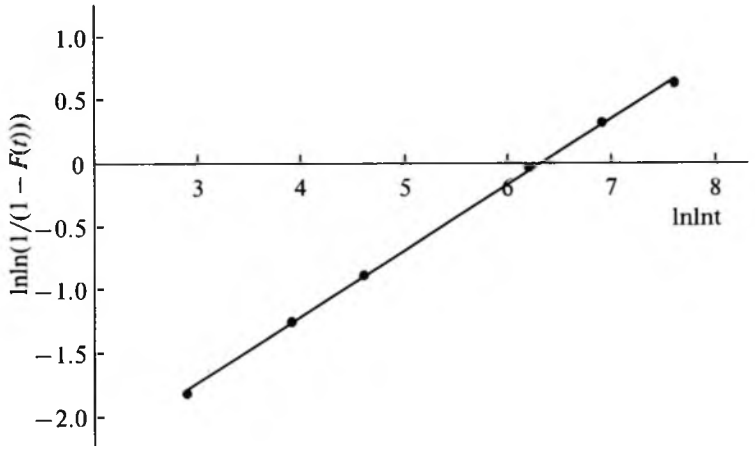


Fig. 4. Weibull's linear scattering of journal publications.

for the approximate calculation of the boundaries of the core and zones of scattering. This method is much simpler compared with the analytical method as it does not require any calculation of distribution.

In the case of homogeneous samples, some rank-size statistical distributions can be described by Weibull's law, (13) and (14). The estimated parameters of this law can be most easily calculated by using the least squares method. If the rank-size distribution contains a heterogeneous part, for example, function words, the third parameter δ needs to be calculated based on formula (30) according to the known values of the α and β parameters, as well as the relative frequency of the first word.

Our study has demonstrated the high accuracy of the approximation of some statistical rank-size distributions using Weibull's law. However, it is recommended to use the second system of continuous distributions and general stable method for a guaranteed calculation of the best theoretical rank distribution by statistical series.

REFERENCES

1. Mikhailov, A.I., Chernyi, A.I., and Gilyarevskii, R.S., *Osnovy informatiki* (Fundamentals of Informatics), Moscow: Nauka, 1968.
2. Neshitoi, V.V., Universal laws of spreading and ageing of publications, *Vesn. Belarus. Dzyarzh. Univ. Kul'tury Mast.*, 2007, no. 8, pp. 128–133.
3. Neshitoi, V.V., *Elementy teorii obobshchennykh raspredelenii* (Elements of Generalized Distribution Theory), Minsk: RIVSh, 2009.
4. Neshitoi, V.V., *Matematiko-statisticheskie metody analiza v bibliotechno-informatsionnoi deyatel'nosti: ucheb.-metod. posobie* (Mathematical-Statistical Methods of Analysis in Librarian-Information Activity: A Tutorial), Minsk: Bel. Gos. Univ. Kul'tury Iskusstv, 2009.
5. Neshitoi, V.V., *Metody statanaliza v bibliotechnoi deyatel'nosti: vychislenie nepreryvnykh raspredelenii: ucheb.-metod. posobie* (Methods of Statistical Analysis in Librarian Activity: A Tutorial), Minsk: Bel. Gos. Univ. Kul'tury Iskusstv, 2010.
6. Neshitoi, V.V., Zipf's and Bradford's laws and universal models, *Autom. Docum. Mathem. Linguist.*, 2010, vol. 44, np. 1, pp. 30–37.
7. Neshitoi, V.V., Form of range distribution presentations, *Uch. Zap. Tartusk. Gos. Univ.*, 1987, no. 774, pp. 123–134.
8. Belonogov, G.G., On some statistical regularities in Russian written speech, *Vopr. Yazykozn.*, 1962, no. 1, pp. 100–101.
9. Neshitoi, V.V., Laws of word distribution in text and its lexical parametrization, *Candidate. Sci. (Philol.) Dissertation*, Minsk, 1973.
10. Lyashevskaya, O.N. and Sharov, S.A., Frequency Dictionary of Contemporary Russian Language (on the Materials of National Corps of Russian Language). http://dict.ruslang.ru/freq.php

Translated by V. Kupriyanova-Ashina

SPELL: 1. unimodal