Scientometrics, Vol. 5. No. 6. (1983) 375–395

THE "RANK DISTORTION" EFFECT AND NON-GAUSSIAN NATURE OF SCIENTIFIC ACTIVITIES

S. D. HAITUN

Institute of History of Science and Technology, Academy of Sciences of the USSR, Staropansky per., 1/5, Moscow (USSR)

(Received March 1, 1983)

The "rank distortion" of statistical distribution and its effect on the non-Gaussian nature of scientific activities is discussed. Examples are presented and in particular, the dispersion of publications by journals (the Bradford distribution) is discussed in detail. The data supporting the thesis of non-Gaussian nature of science are reexamined, and the empirical basis of the thesis is extended.

Introduction

In a recent publication[1] the thesis of non-Gaussian nature of scientific activities was formulated and extended to human activities in general. This empirical observation has rather profound theoretical consequences. The most direct of these consequences is the need to put the quantitative analysis of scientific activities (measurements, mathematical modelling, decision-making) on the rails of non-Gaussian statistics. This and other less evident but not less important consequences form the subject matter for a series of subsequent publications. But, from the very outset the need to reject the traditional Gaussian mathematical statistics, i.e. the apparatus of moments: mean, dispersion, Pearson's correlation coefficient, factor analysis, the method of least squares, etc. instils the author with some fear. Therefore, before turning to decisive reforms in the quantitative analysis of scientific (and, in general, human) activities, I feel it imperative to discuss certain details of empirical principles of the above-mentioned thesis. First, some errors committed previously in handling the "rank distortion" effect[2] are to be eliminated. Second, the empirical basis will be extended.

How the non-Gaussian nature of scientific activities was detected

The non-Gaussian nature of scientific activities was inferred from an analysis of empirical stationary distributions comparising samples from 47 scientometric and 58 non-scientometric distributions. This series of samples was compiled without any specific intention. Hence it may be regarded as representative to stationary distributions of scientific and human activities in general. Non-Gaussian nature of the majority of the distributions was established.

We consider *non-Gaussian* the statistical distributions for which the central limit theorem is not satisfied, but the Gnedenko-Doeblin limit theorem holds. These theorems predict the behaviour of the distribution of the sum of identically distributed independent random variables when the number of terms in the sum tends to *infinity.* The nature of the dependence of the moments on the sample size. Infinity is however a mathematical absraction. The dependence of distributions moments on the sample size offers a practical criterion for testing whether a distribution is Gaussian or non-Gaussian. If this dependence is significant (for the problem under consideration), the distribution is non-Gaussian; if insignificant, the distribution is Gaussian.

In the analysis of non-Gaussian distributions a key role is played by the *Zipf distribution.* Namely, according to the Gnedenko–Doeblin theorem for large values of the random variable non-Gaussian distribution take the form of Zipf distribution up to some "slowly changing" function. The frequency form of the Zipf distribution is

$$n(x) = \frac{C}{x^{1+\alpha}}, \quad 0 < \alpha < \infty, \quad x \geqslant x_0 > 0, \tag{1}$$

where x is the random variable, n(x) is the frequency, C is a parameter depending on the sample size (see Eq. (14)); the exponent α characterizes the "degree of non-Gaussian nature" of the Zipf distribution (as α decreases, the non-Gaussian nature increases). If we adhere to a rigorous mathematical approach, i.e. if we are dealing with the behaviour of the distribution of the sum of independent identically distributed random variables (obeying the Zipf law) with the number of terms in this sum tending to *infinity,* a Zipf distribution is Gaussian if $\alpha > 2$, and non-Gaussian if $\alpha \leqslant 2$. But, in practice, a Zipf distribution with $\alpha > 2$ (as estimated from a finite sample) has often to be considered non-Gaussian. This happens particularly, if the moments depend essentially on the sample size. For a fixed α, the smaller the sample size, the more significant becomes the dependence of the moments of the Zipf distribution on the sample size. Therefore in judging whether a given empirical distribution is Gaussian or non-Gaussian, one has to take into account not only the value of α, but also the sample size.

To estimate the dependence of the moments on the sample size, empirical data were approximated by a Zipfian distribution[3] and α was estimated. To this end, empirical data were plotted on double logarithmic paper, an asymptote was drawn for large x, and the value of α was estimated by the slope of the asymptote. Comparing the values of α and the sample size with the graphs regarding the dependence of the moments of the Zipf distribution on the sample size (more precisely, on the maximum value of the variable J), the degree of dependence of the moments on the sample size could be judged.

Is the use of the Zipfian approximation justified in general? Strange as it is, such an approximation also includes Gaussian non-Zipfian distributions, namely, Gauss, Poisson, lognormal, negative binomial, and other distributions. This is strange, because for all these distributions α turns out to be infinity, i.e. the Zipfian approximation with an $\alpha < \infty$ is, apparenthy, inapplicable to them. However, it should be taken into consideration that for a Gaussian non-Zipfian distribution $\alpha = \infty$ prevails only for an *infinite* sample size. But, since an empirical sample is always finite, the limiting value $\alpha = \infty$ is practically never reached for Gaussian non-Zipfian distributions either. Therefore the Zipfian approximation proves to be justified in the general case of stationary empirical distribution, both Gaussian and non-Gaussian.

Having found the value of α for each of the 105 distributions, we plotted the rank distribution of these values on double logarithmic paper (Fig. 1)[4]. This diagram shows that a major part of the empirical distributions had low values of α. Since, as a rule, sample sizes were small, the dependence of moments on sample size appeared to be significant, indicating the non-Gaussian nature of the majority of the distributions in question.

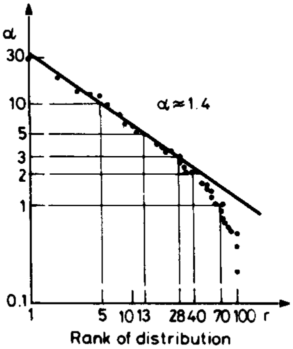

Fig. 1. Rank distribution of the exponent α of the Zipfian distribution in 105 empirical stationary distributions[5] ("rank distortion" effect discarded)

Many scientometric and non-scientometric distributions of human activity have to be taken in a *rank* form. This is so because as shown earlier[1] for small sample sizes the frequency form of distributions cannot be applied, and the rank form is to be used instead. The situation is all the more complicated by the "rank distortion" effect: in rank representation in the range of high values of the variable (i.e. just in the region where the value of α is found graphically), the graph of Zipfian distribution may deviate from a straight line (on double logarithmic scale) for certain relationships between the sample parameters. In the mentioned paper[1] this effect was taken into account only partially. Namely, it was shown that overlooking this effect leads to overestimation of the graphically estimated value of α. Therefore, were the "rank distortion" effect, as has been said, is taken into account, it might only *strengthen* the thesis about non-Gaussian nature of scientific activities.

In what follows the "rank distortion" effect is analyzed more rigorously in connection with the problem of non-Gaussian nature of scientific activities.

The theory of "rank distortion" effect

Any *discrete* statistical distribution has, besides the usual *frequency* form, also a *rank* form. *The frequency differential form* is determined by n(x), i.e. the frequency of occurence of a given value of the random variable x in a sample of size N:

$$\sum_{x_0}^{J} n(x) = N, \tag{2}$$

where x_0 and J are the sample minimum and maximum values of x, respectively.

The frequency integral form is given by

$$F(x) = \frac{1}{N} \sum_{x_0}^{x} n(\xi), \tag{3}$$

where F(x) is the distribution function.

The rank form is introduced by the relation

$$r = \sum_{x}^{J} n(\xi). \tag{4}$$

The rank r has a simple meaning: this is the ordinal number of a given value of the random variable when all these values are arranged in a decreasing order. Thus, for a sample of size N we have N ranks

$$1 \leqslant r \leqslant N. \tag{5}$$

Different ranks are assigned to *equal* values of x.

The rank differential form is determined by x(r), which is found in an explicit form from Eq. (4).

The rank integral form is given by

$$X(r) = \sum_{1}^{r} x(\xi), \quad \sum_{1}^{N} x(\xi) = G. \tag{6}$$

As has been mentioned, the definition Eq. (4) is introduced for a *discrete* distribution. For a *continuous* distribution neither the definition[6]

$$r = N[1 - F(x)], \quad 0 \leqslant r \leqslant N, \tag{7}$$

nor the definition[7]

$$r = 1 + N[1 - F(x)], \quad 1 \leqslant r \leqslant N + 1, \tag{8}$$

only their combination

$$r = \left| \begin{array}{ll} N[1 - F(x)], & r \geqslant 1, \\ & \\ 1 + N[1 - F(x)], & r \leqslant N, \end{array} \right. \tag{9}$$

is correct, which provides a necessary range of r i.e. from 1 to N. This is so due to the difference between an integral and a sum:

$$\sum_{x}^{J} n(\xi)\,|_{x=y} \equiv 1; \quad \int_{x}^{J} n(z)d\xi\,|_{x=J} \equiv 0. \tag{10}$$

The rank differential form of the Zipf distribution is:

$$x(r) = \frac{A}{(r + B)^{\gamma}} \tag{11}$$

The rank integral form of the Zipf distribution is given by:

$$X(r) = \left| \begin{array}{ll} A \ln \dfrac{r + B}{1 + B}, & \gamma = 1 \\ & \\ \dfrac{A}{\gamma - 1} [(1 + B)^{1-\gamma} - (r + B)^{1-\gamma}, & \gamma \neq 1. \end{array} \right. \tag{12}$$

In formulae (1) and (11)–(12).

$$y = \frac{A}{(1+B)^{\gamma}}, \quad x_0 = \frac{A}{(N+B)^{\gamma}} \tag{13}$$

and

$$\gamma = \frac{1}{\alpha}; \qquad A = \left[\frac{N-1}{\frac{1}{x_0^{\alpha}} - \frac{1}{J^{\alpha}}}\right]^{\frac{1}{\alpha}};$$

$$B = \frac{N-1}{\left(\frac{J}{x_0}\right)^{\alpha} - 1} - 1; \quad C = \frac{\alpha(N-1)}{\frac{1}{x_0^{\alpha}} - \frac{1}{J^{\alpha}}}. \tag{14}$$

The "rank distortion" effect, i.e. deviation of the Zipfian distribution x(r) from a straight line for large values of x (small values of r) on double logarithmic scale, occurs owing to the presence of the parameter B in the denominator of the expression (11). For this reason we refer to B as *rank distortion coefficient.* This effect is significant, evidently, if $B \gtrsim 1$. For a Zipf distribution, B being given by Eq. (14), the condition for the "rank distortion" effect to be significant takes the form

$$\frac{y}{x_0} \lesssim \left[\frac{N+1}{2}\right]^{\frac{1}{\alpha}} \qquad 15)$$

For a general Zipfian distribution the condition Eq. (15) remains the same, except for the difference that N is to be replaced by the maximum rank at which a given Zipfian distribution retains the Zipf distribution form, and x_0 by an x that corresponds to this maximum rank.

Thus, the "rank distortion" does not always manifests itself, but only under an "unfavourable" combination of values of the sample parameters N, x_0 and J and the parameter of the Zipfian distribution α. At fixed N and x_0 the less are α and J, the more significant is this effect. As we see, all other conditions being equal, the more non-Gaussian is a given empirical distribution, the more significant is the "rank distortion" effect.

Examples of "rank distortion" effect

The "rank distortion" effect complicates the approximation of empirical data by Zipfian distribution with a definite value of α. Sometimes it is difficult to give preference to one or other approximation, difficult to judge whether we are dealing with a Zipf distribution with one value of a or with a Zipfian distribution, which is not a Zipf distribution, with a higher value of α. The situation is still more complicated because the conventional mathematical estimation techniques, such as, the method of least squares cannot be applied here due to the non-Gaussian nature of the distributions. The non-Gaussian estimation methods (including, probably, the maximum likehood method and the chi-square method in its parametric version) have not yet been elaborated or developed for non-Gaussian distributions. All this is still to be done, but at present we have to restrict ourselves to graphical fitting of empirical data, so to say, by a rule of thumb.

In this section, we present three examples of empirical distributions where the "rank distrotion" effect is significant. One more example, the Bradford distribution, is taken up in a separate section as it plays a central role in scientometrics and innumerable papers are devoted to it.

Example 1. Distribution of laboratories by the number of references to their papers (Fig. 2 a, b). In the paper mentioned above[1] using the rank representation (Fig. 2a) and disregarding the "rank distortion", a Zipfian distribution was fitted to the data with $\alpha = 3.5$. However, the data also admit fitting a Zipf distribution with $\alpha = 1.0$. The dashed line in Fig. 2a corresponds to the Zipf distribution with sample parameters $N = 34$, $x_0 =$ $= 5$, $J = 31$, and $\alpha = 1.0$. The expression for this approximation was determined as fol-

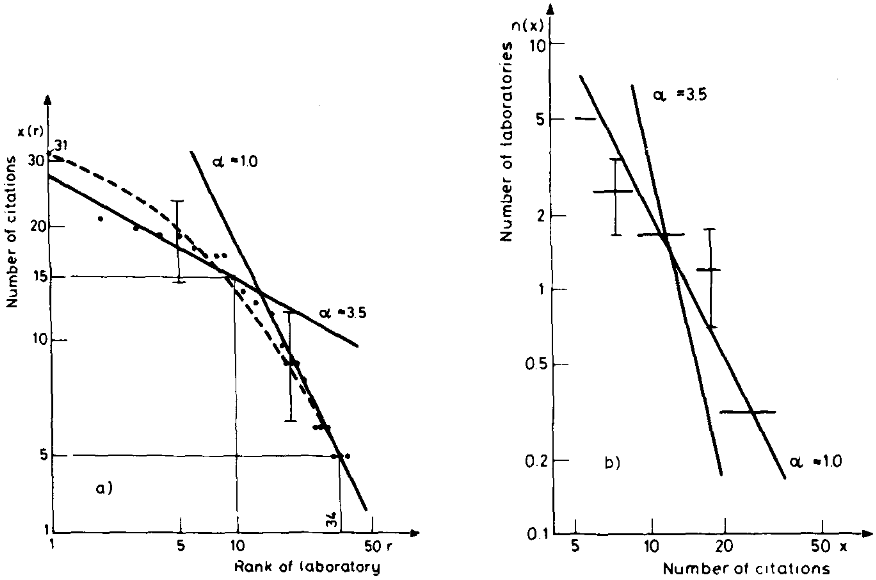

Fig. 2. Distribution of ceramic materials research laboratories by the number of citations to their publications (data from Ref.)[8]

lows. Substituting the values of the parameters into Eq. (14), we find B = 5.35. Thus. using Eqs (11) and (13), we obtain the expression for the unknown Zipf distribution

$$x(r) = J\left(\frac{1+B}{r+B}\right)^{\frac{1}{\alpha}} = 34 \cdot \left(\frac{1+1.35}{r+5.35}\right)^{1.0} .$$

As is seen, the Zipf distribution describes quite satisfactorily the empirical rank data. The frequency curve (Fig. 2b) compels us to prefer the second of the above-mentioned alternatives ($\alpha = 1.0$).

Example 2. Distribution of biological genera by the number of species (Fig. 3a, b). This is the well-known Willis distribution which is usually regarded as biological according to the nature of the classified objects. In our opinion, it is an improper practice to identify taxonometric distributions by the objects of classification. These kinds of distributions gradually *lose their form* as the accuracy of determination of classification attributes increases. In the limit when even the slightest differences, could be detected each object would be classified in its "personal" group. As the measurement error increases, more and more objects get under each classification group and the distribution acquires a definite form. Thus, classification distribution acquires a distinct shape as the accuracy decreases, and loses it is the accuracy increases. Usually, the radiobility of description of objects increases *with increasing measuring accuracy.* If at all the raliability of the description increases with decreasing observation accuracy, it comes from the Evil or the observer. We have come to the conclusion that a classification distribution, while describing implicitly the classification principle, reflects the "human" nature of the investigator elaborating a concrete classification scheme, and is, like any other distribution of human activity, of the Zipfian form. Such are, e.q., the distributions of classes of inorganic minerals by the number of groups in a class, the distributions of sections of the universal decimal classification (adopted in the USSR libraries) by the number of subsections in different branches of scientific disciplines (physics, chemistry, zoology, biology, management theory, etc) and many others. One can even speculate that the more creative is the investigator, the more non-Gaussian is the classification developed by him, i.e. the smaller is the value of α characterizing the corresponding taxonometric distribution (see p. 392).

On the basis of the frequency form of the Willis distribution representation (Fig. 3a) it is usually described by the Zipf distribution with $\alpha = 0.5$. The frequency data do indeed admit such a fit. Let us, however, turn our attention to the rank representation (Fig. 3b). Here the dashed line is a Zipf distribution with $\alpha = 0.5$ and sample parameters N = 200, $x_0 = 1$, J = 106 and B = 20.4:

$$x(r) = 106 \cdot \left(\frac{1+20.4}{r+20.4}\right)^{\frac{1}{0.5}}$$

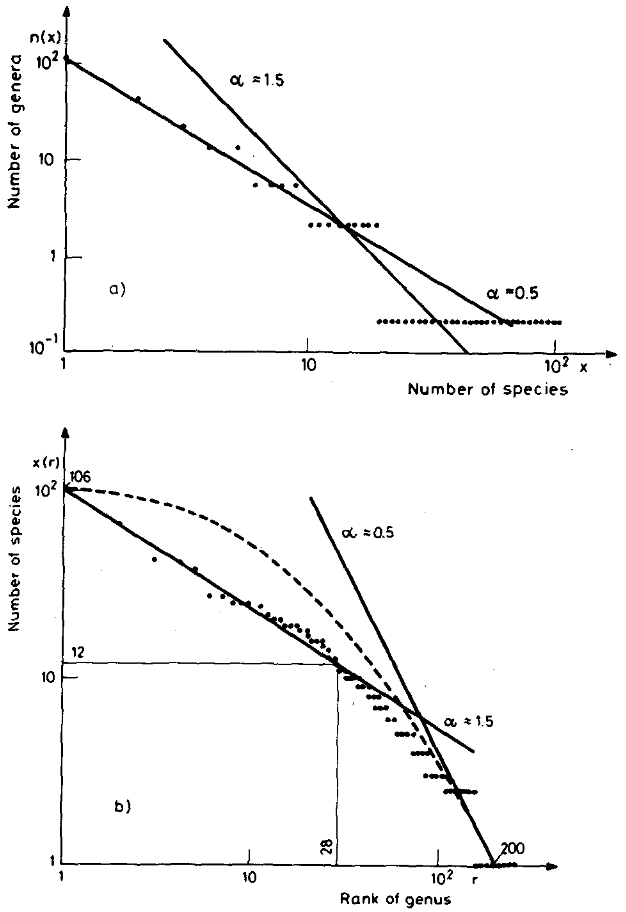

Fig. 3. Taxonomic distribution of biological genera by the number of species for lizards (data of Willis[9]).
(a) Frequency form.
(b) Rank form.

It is seen that the empirical data are rather poorly described by the Zipf distribution with $\alpha = 0.5$. Therefore, we shall dwell on a Zipfian distribution with $\alpha = 1.5$, that describes satisfactorily the data both in the rank and the frequency representations.

Example 3. Distribution of scientists by partial productivity (Fig. 4a, b). Based only on the rank representation (Fig. 4a) and disregarding the "rank distortion", a Zipfian distribution with $\alpha = 4.4$[10] has been fitted to the data, which is not a Zipf distribution.

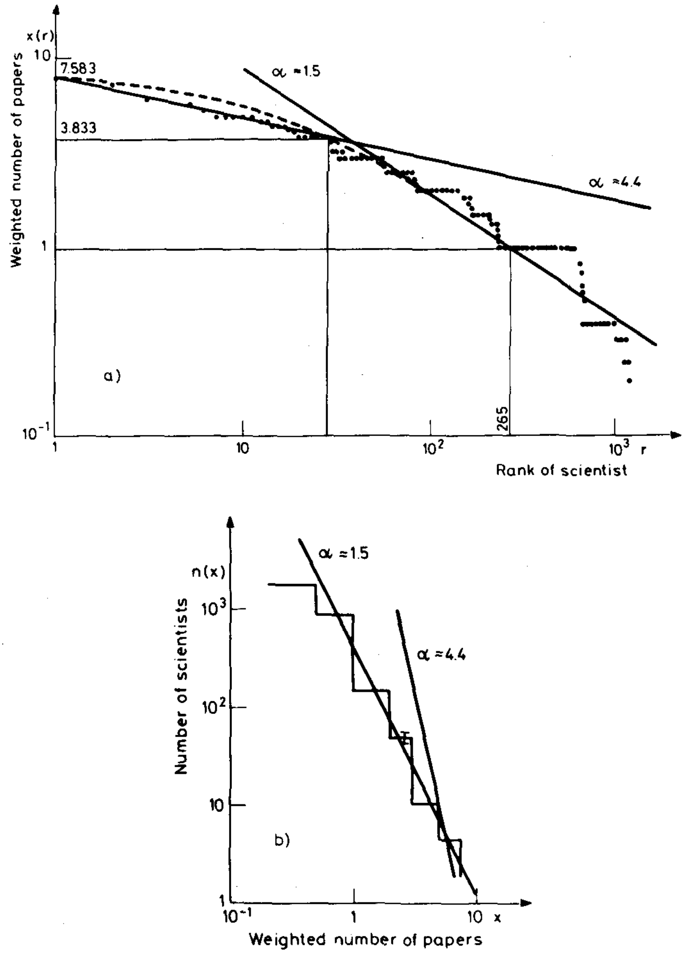

Fig. 4. Distribution of scientists by partial productivity (for a paper written by n authors, each author is credited 1/n authorships). (Section "Thermodynamics and statistical physics" of Referativnij Zhurnal, Fizika, 1975, 1976 and the first half of 1977)

Indeed, rank data well admit such a fit. However, it appears (Fig. 4b) that a much better fit to the frequency data is a Zipfian distribution with $\alpha = 1.5$, which is not a Zipf distribution either. Going back to the rank representation (Fig. 4a), we find that this second approximation also describes satisfactorily the data in the rank form. In the diagram the dashed curve corresponds to a Zipfian distribution with $\alpha = 1.5$, $N = 265$, $x_0 = 1$, $J =$ $= 583$ and $B = \{264 / [(7.583/1)^{1.5} - 1]\} - 1 = 12.3$:

$$x(r) = 7.583 \left(\frac{1 + 12.3}{r + 12.3}\right)^{\frac{1}{1.5}}$$

This approximation describes, satisfactorily though not perfectly the rank data. Therefore, we prefer the estimation, $\alpha = 1.5$, which describes better the frequency and the rank data on the whole.

"Rank distortion" and Bradford distribution

Bradford distribution is a term applied to the distribution of papers on a particular topic by journals. There exists a large number of approximations for this empirical distribution (see, for example, Ref.[11]. We shall here show that lack of clarity on this question owes its origin to the unsuccessful choice of the integral rank form made by Bradford for representing statistical distribution of data. The use of the frequency differential form or, if the sample size is not sufficiently large, the rank differential form mitigates the situation: the Bradford distribution would be approximated by the Zipf distribution, provided the "rank distortion" is large.

From the very start, *Bradford*[12] built on the integral rank form of data representation and proposed the approximation

$$X(r) = a + b \log r, \tag{16}$$

where r is the rank of the journal, X(r) is the commulative number of papers, a and b are parameters. *Bradford* has chosen the coordinates in such a way that the approximation have the form of a straight line: r was plotted on the abscissa in a logarithmic scale and X(r). on the ordinate in a linear scale (Fig. 5).

The Bradford approximation corresponds to a Zipf distribution with $\alpha = 1$ if "rank distortion" is small. Indeed, a Zipf distribution in its rank integral form for $\alpha = 1$ (see Eq. (12)) is of the form

$$X(r) = A \ln \frac{r + B}{1 + B} = - A \ln (1 + B) + \frac{A}{\log e} \log (r + B).$$

For $r \gg B$ this gives

$$X(r) \cong - A \ln (1 + B) + \frac{A}{\log e} \log r,$$

which corresponds to the Bradford approximation (Eq. (16)) with

$$a = - A \ln (1 + B); \quad b = \frac{A}{\log e}$$

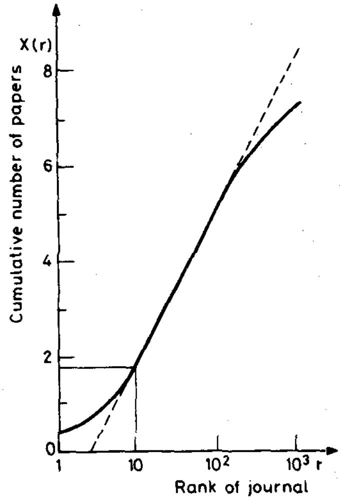

Fig. 5. Bradford's dispersion law.

For $\alpha = 1$, $x_0 = 1$, $N \gg 1$ and $J \gg 1$, as we most often encounter in applications, we have (see Eq. (14)):

$$A = \frac{N-1}{1-\frac{1}{J}} \approx N; \quad B = \frac{N-1}{J-1} - 1 \approx \frac{N}{J} - 1,$$

so that

$$a \approx -N \ln \frac{N}{J}; \quad b \approx \frac{N}{\log e}. \tag{17}$$

Thus, *the parameters of the Bradford approximation are determined by the sample parameters and do not characterize the dispersion of papers by journals,* which is determined, like for any stationary scientometric, i.e. non-Gaussian, distribution, by the value of α. In the Bradford approximation, however, α is fixed and is taken equal to 1.

There are various approximations for empirical data taken in Bradford representation that are used in a situation which the Bradford approximation is to able to cope with and which is most often encountered in applications. Namely, when the graph in Fig. 5 is quite different from a straight line in the region of small r, i.e. when the "rank distortion" is large. *Leimkuhler*[13] uses the approximation

$$X(r) = \frac{a \log(1 + br)}{\log(1 + b)}, \tag{18}$$

which corresponds to a Zipf distribution with $\alpha = 1$ and a large rank distortion, $B \gg 1$. Indeed, the Zipf distribution for $\alpha = 1$ has the form (see Eq. (12)):

$$X(r) = - A \ln(1 + B) + A \ln(r + B).$$

This corresponds to the Leimkuhler approximation

$$X(r) = \frac{a \ln b}{\ln(1 + b)} + \frac{a}{\ln(1 + b)} \ln\left(r + \frac{1}{b}\right)$$

with

$$b = \frac{1}{B}$$

provided the two equations for the Leimkuhler approximation parameter a are consistent:

$$\frac{a}{\ln(1 + b)} = A$$

and

$$\frac{a \ln b}{\ln(1 + b)} = - A \ln(1 + B).$$

The first gives

$$a = A \ln \frac{1 + B}{B}.$$

the second

$$a = A \left(\ln \frac{1 + B}{B}\right) \frac{\ln(1 + B)}{\ln B}.$$

These expressions are consistentt if $B \gg 1$, which was required to be proved.

Brookes[14] approximates the beginning of the graph shown in Fig. 5 by an exponential curve, and the remaining part by a straight line

$$X(r) = \begin{matrix} \delta r^{\beta}, & 1 \leqslant r \leqslant c \\ N \ln \frac{r}{s}, & c \leqslant r \leqslant N, \end{matrix} \tag{19}$$

where δ, β, s, c are parameters. This approximation is a combination of two Zipf distributions with different values of α. The first part of the Brookes approximation

$$x(r) = \frac{dX(r)}{dr} = \frac{\delta\beta}{r^{1-\beta}}$$

corresponds to the Zipf distribution with $\alpha = 1 / (1 - \beta)$ and the rank distortion coefficient $B = 0$ (see Eq. (11)). The secon part is the Zipf distribution with $\alpha = 1$ and $B = s - 1 \ll r$ (see Eq. (12)).

It has so far not been possible to describe the curve in Fig. 5 satisfactorily by formulae. Precisely for this reason, there exist different approximations of this diagram. In particular, it is rather difficult to explain why the *upper* part of the empirical curve is different from a straight line (the so-called "droop" or the effect of *Groos*[15] who first noticed this anomaly). For this reason, side by side with the *integral* rank form the publication dispersion curves are also represented in some papers in a more traditional differential frequency or rank form. In this case the Bradford distribution is associated with the Zipf distribution. Such approximations has been dealt with earlier by the author[16].

Although there is no satisfactory approximation of the publication dispersion curve in the Bradford representation, the latter is so far encountered often in applications. This happens because the *rank* representation is mainly used in applications, and the Bradford representation is a modified rank form. Moreover, it is well known that the *differential* rank form on double logarithmic scale does not give a straight line which is believed to correspond in this representation to the Zipf distribution

$$x(r) \sim \frac{1}{r^{\gamma}} \tag{20}$$

This is evidently the reason why Bradford proposed his rather refined representation.

We now know that Eq. (20) describes the Zipf distribution only in a particular case where the "rank distortion" is negligibly small. For a general Zipf rank distribution the "Mandelbrot law" Eq. (11) is valid. Empirical data on scatter of publications by journals are well described by the Zipf distribution both in the frequency and in the rank form. We shall illustrate this on an example of typical empirical data of this kind (see Fig. 6 a–c). Figure 6a, which shows the data in Bradford representation, provides evidences that these data are really typical. There is no Groos effect (the upper part of the graph is here on the straight line), possibly, because the number of journals in this case is not sufficiently large. Figure 6b shows data in a frequency differential form, and Fig. 6c in rank differential form. In this latest diagram the dashed line corresponds to a Zipf distribution with sample parameters $N = 1100$, $x_0 = 1$, $J = 36$ and $\alpha = 0.80$. The rank distortion coefficient B, according to Eq. (14) is 65.3. Therefore, the expression for the Zipf distribution takes the form

$$x(r) = y\left(\frac{1+B}{r+B}\right)^{\frac{1}{\alpha}} = 36 \cdot \left(\frac{1+65.3}{r+65.3}\right)^{\frac{1}{0.80}}$$

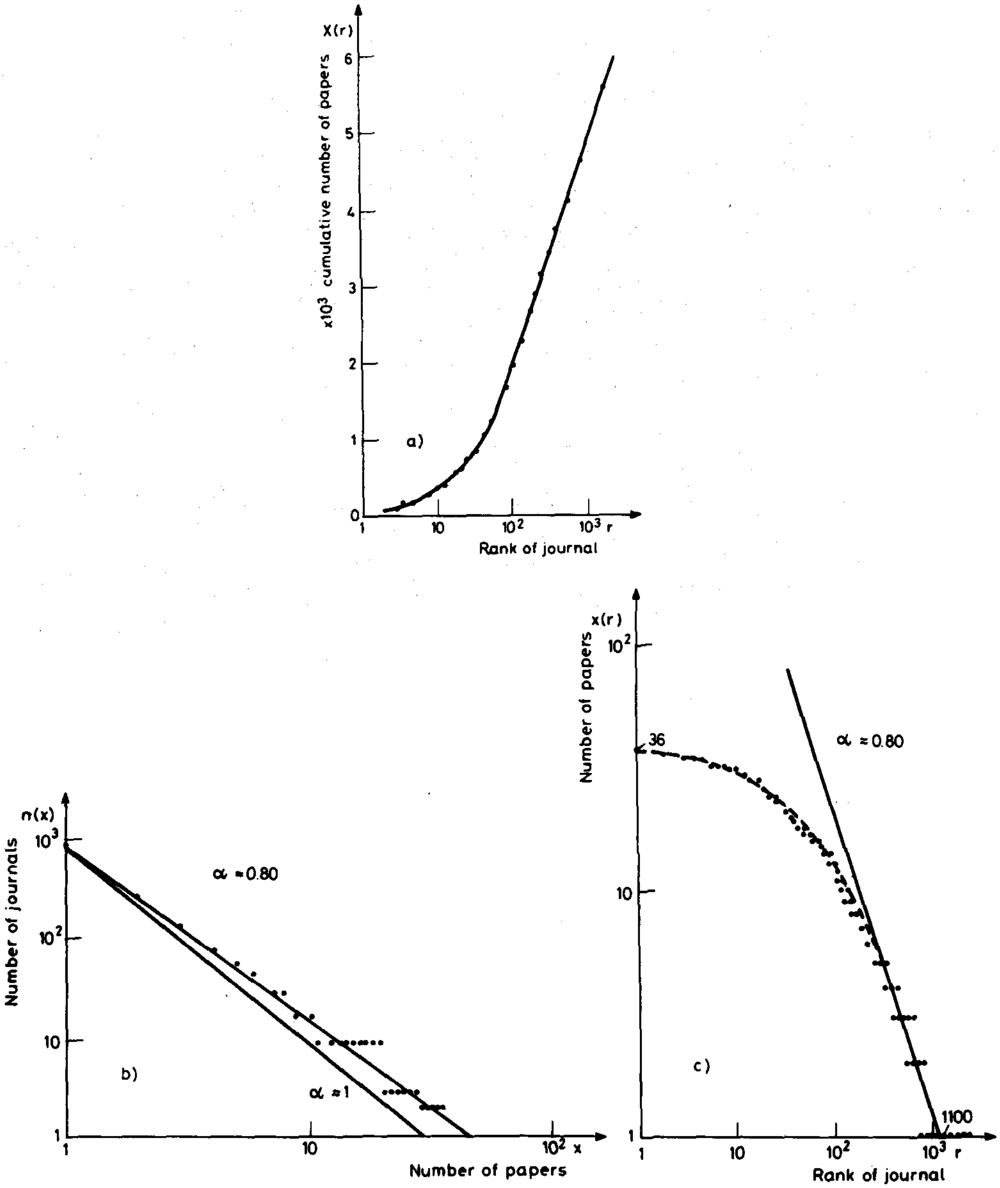

Fig. 6. Distribution of journals by the number of papers on a given topic (medicine) (data from Ref.)[17]

(a) The Bradford representation.

(b) Frequency form.

(c) Rank form.

As we see, the data admit approximation by the Zipf distribution with $\alpha = 0.8$.

Summing up, one can say that in general the Bradford distribution is Zipfian with an α approximately equal to 0.8–1.0. Deviation of the Bradford distribution from the Zipf distribution for small x (large r) may, probably, be identified with the Groos droop. This deviation appears for a large sample size of journals. For a smaller sample size the Bradford distribution may quite well be approximated by the Zipf distribution. If the Groos droop is perceptible, one should use other Zipfian approximations, possibly, Price's beta-function.

The Bradford representation, often used now in applications, is less convenient than representation in the differential form in log-log coordinates. Moreover, the Bradford distribution is, to put it bluntly, *incorrect,* since the straight line in it corresponds to a Zipf distribution with a fixed $\alpha = 1$. But α may vary from sample to sample and characterize, properly speaking, the degree of publication dispersion by journals. The slope of the linear region of the Bradford approximation (Figs. 5, 6a) $b \approx N / \log e$ (see Eqs 16 and 17) does not, therefore, characterize the degree of publication dispersion, as sometimes believed, but solely the number of journals in a given sample. For all these reasons the Bradford representation should give way to the differential (rank and frequency) data representation in log-log coordinates.

Non-Gaussian nature of scientific activity: new data

Empirical observations of the non-Gaussian nature of scientific activity were based on a series of samples from 105 stationary distributions.[1] Those given in the rank form were reanalyzed and the values of α were reestimated taking the "rank distortion" into account. In those cases where it was impossible to give preference to one or other estimation, the one with a higher α was chosen, thus, weakening the thesis of non-Gaussian nature of scientific activity. The results are presented in Fig. 7. As supposed, taking "rank distortion" into account, did not disprove the thesis under discussion, but, on the contrary, has given greater support in favour of it. But it did not strengthen, as has been expected, the thesis much. This is expressed in the fact that the distribution of the values of α has now a somewhat lower $\alpha = 1.23$, in place of 1.4. And $\alpha \leqslant 1$ is observed not in 33% cases, as before, but already in 41%, and the corresponding per cent of distributions with $\alpha \leqslant 2$ has changed from 62 to 63. Thus, a still larger part of our 105 distributions, than was determined without regard for the "rank distortion", has moments which essentially depend on the sample size. This just means strengthening of the empirical basis of the thesis concerning non-Gaussian nature of scientific, and other human activities.

In order to strengthen further this empirical basis, we have extended also the sample of stationary distributions, by adding to the samples 45 scientometric distributions,[19] and

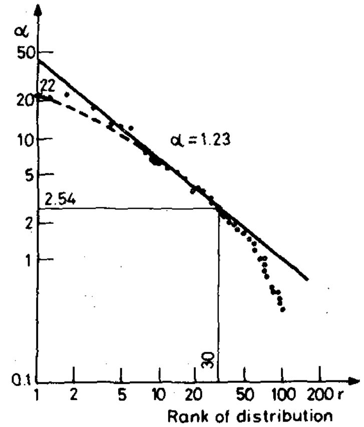

Fig. 7. Rank distribution of the exponent α of the Zipfian distribution in 105 empirical stationary distributions

20 non-scientometric distributions.[20] These new 65 distributions were analyzed similarly to the previous 105 ones. The results are presented in Fig. 8. As is seen, extension of empirical basis leads to the old results, namely, the moments of the majority of the extended set of 170 stationary distributions essentially depend on the sample size.

It has been found that the stationary distributions of scientific activities have on the whole a lesser α than that of other forms of human activities[21]. We present in confirmation

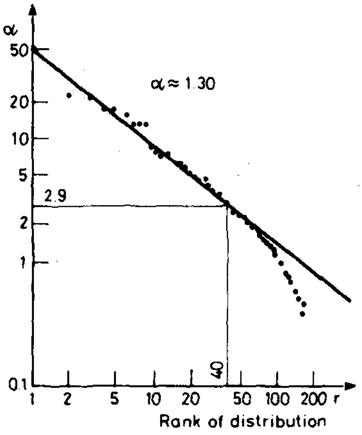

Fig. 8. Rank distribution of α in 170 empirical stationary distributions

the data on the extended set of samples. Figure 9 illustrates distributions of α for 92 scientometric, and Fig. 10 – for 78 non-scientometric social distributions. Comparing Figs. 8, 9 and 10 we see that, indeed, to the extent to which our samples are representative, scientific activities are more non-Gaussian than other types of human activities and than human activities as a whole. This may be due to the *creative* nature of scientific activities. Probably, the more creative is a given human activity, the more non-Gaussian it is. This is equally true for individuals, groups of individuals, as well as for scientific and other social communities.

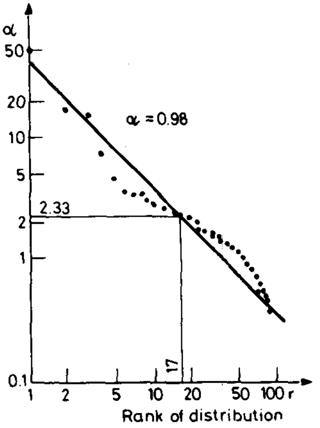

Fig. 9. Rank distribution of α in 92 scientometric stationary distributions

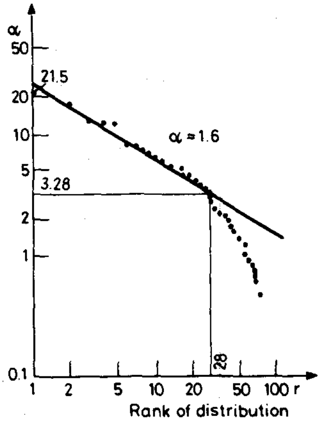

Fig. 10. Rank distribution of α in 78 social non-scientometric distributions

In conclusion, we note that the program outlined in the introduction is fulfilled. The empirical basis of the thesis about non-Gaussian nature of scientific activities has been given greater clarity by taking the "rank distortion" into consideration, and the sample size on which the thesis is based has somewhat been extended.

Notes and References

1. S. D. HAITUN, Stationary scientometric distributions. Part 1. The different approximations, *Scientometrics,* 4 (1982) 5–25; Part II. Non-Gaussian nature of scientific activities. – *Scientometrics* 4 (1982) 89–104; Part III. The role of the Zipf distribution. – *Scientometrics* 4 (1982) 181–194.
2. There are some important inaccuracies in Ref. 1, which are not touched upon in the main text of the present paper.

   The conditions of applicability of the frequency form of statistical distribution (formulae (A.6) and (A.7) of Part I) should be written as follows

   $$\frac{\sigma_n}{n} \approx \frac{\sqrt{n}}{n} = \frac{1}{\sqrt{n}} \ll 1;$$

   $$\frac{J - x_0}{\Delta x} \ll N.$$

   The condition of applicability of the rank form of statistical distribution (formula (A.9) of Part I) should be written as

   $$\frac{J + x_0}{\Delta x} \gg 2.$$

3. We call the distribution Zipfian if at high values of the variable it has the form of the Zipf distribution.
4. In Ref. 1 the value of α corresponding to the upper mark on the ordinate was erroneous.
5. S. D. HAITUN, op. cit., Note 1, Part II.
6. See, for example, A. I. YABLONSKY, Stokhasticheskiye modeli nauchnoi deyatelnosti (Stochastic models of research activities). in: *Sistemniye issledovaniya, Yezhegodnik,* 1975. (*System Research, Yearbook,* 1975), Nauka, Moscow, 1976, p. 5–42; A. I. YABLONSKY, On fundamental regularities of the distribution of scientific productivity, *Scientometrics* 2 (1980) 3–34.
7. This error is made by the present author in Ref. 1, see S. D. HAITUN, op. cit., Note 1, part I.
8. J. H. WESTBROOK, Identifying significant research, *Science,* 132 (1960) 1229–1234.
9. B. HILL, Zipf's law and prior distribution for the composition of a population, *J. Am. Stat. Assoc.,* 65 (1970) 1220–1232.
10. See S. D. HAITUN, op. cit., note 1, part II, Fig. 2.5.
11. S. D. HAITUN, op. cit., note 1, part I.
12. S. C. BRADFORD, Sources of information on specific subjects, *Engineering,* 26 (1934) January; S. C. Bradford Documentation, London, Grosby Lookwood and Son Ltd, 1948.
13. F. F. LEIMKUHLER, The Bradford distribution, *J. Docum.,* 23 (1967) 197–207; F. F. LEIMKUHLER, Operational analysis of library systems, *Information and Management,* 13 (1977) 79–93.

14. B. C. BROOKES, The derivation and application of the Bradford–Zipf distribution, *J. Docum.*, 24 (1968) 247–265; B. C. BROOKES, Bradford's law and the bibliography of science, *Nature*, 224 (1969) 953–956; B. C. BROOKES. Theory of the Bradford law, *J. Docum.*, 33 (1977) 180–209.
15. O. V. GROOS, Bradford's law and the Keenan–Atherton data, *Amer. Docum.*, 19 (1967) 46.
16. These approximations are presented in: S. D. HAITUN, op. cit., note 1, part I.
17. W. GOFFMAN, K. S. WARREN, Dispersion of papers among journals based on a mathematical analysis of the diverse medical literatures, *Nature*, 22 (1969) 1205–1207.
18. See S. D. HAITUN, op. cit., note 1, part III, p. 182.
19. S. COLE, J. R. COLE, Visibility and the structural bases of awareness of scientific research, *Am. Sociol. Rev.*, 33 (1968) 397–483; B. V. DEAN, Evaluating Selecting and Controlling R & D Projects, *American Management Association*, Inc. 1968; N. C. MULLINS, The distribution of social and cultural properties in informal communication network among biological scientists, *Am. Sociol. Rev.*, 33 (1968) 781–797; D. CRANE, Social structure in a group of scientists; A test of the "invisible college" hypothesis, *Am. Sociol. Rev.*, 34 (1969) 335–352; W. GOFFMAN, K. S. WARREN, op. cit., note 17; S. CRAWFORD, Informal communication among scientists in sleep research, *Am. Soc. Inform. Sci.*, 22 (1971) 301–310; D. CRANE, *Invisible colleges*, Chicago–London, 1972; R. K. MERTON, *The Sociology of Science. Theoretical and Empirical Investigations*, The Univ. Chicago Press, Chicago, 1973; *Problemy Deyatelnosti uchenogo i nauchnykh kollektivov (Problems of scientists and scientific organization activities)*. Iss. 5, Leningrad, Nauka, 1973; Iss. 6, Moscow, Leningrad, Nauka, 1977; Iss. 7, Moscow–Leningrad, Nauka, 1979; S. S. BLUME, R. SINCLAIR, Aspects of the structure of a scientific discipline, in: Social Processes of Scientific Development. R. WHITLEY Ed. Boston–London, 1974, p. 224–241; *Sotsiologicheskiye problemy nauki (Sociology of science*, Moscow, Nauka, 1974; N. S. ENDLER, J. P. RUSHTON, H. L. ROEDINGER. Productivity and scholary impact (citation) of British, Canadian and U.S. departments of psychology (1975), *Am. Psychol*, 33 (1978) 1064–1082; R. E. EVENSON, Y. KISLEY. *Agricultural research and productivity*, New Hawen, Yale University Press, 1975; K. D. KNORR. The nature of scientific consensus and the case of the social sciences, in: *Determinants and Controls of Scientific Development*, D. Reidel Publ. Co., Dordrecht, Boston, 1975, p. 227–256; *Narodnoye Khozyaystvo SSSR v 1975 g. (Economics of the USSR in 1975)*, Moscow, Statistika, 1976; B. C. FREEMAN, Faculty women in the American university: up the down staircase, *Higher Education*, 6 (1977) 165–188; H. ZUCKERMAN. *Scientific elite: Nobel Laureates in the United States*, Free Press, N.Y, 1977; B. LUBIN, R. G. NATHAN, J. D. MATARAZZO. Psychologists in medical education, 1976, *Am. Psychol.*, 33 (1978) 339–343; Sotsialnoye upravleniye v nauke (Social management in science). Moscow ISI AN SSSR, 1978; H. INHABER, M. S. LIPSETT. Gaps in "Gaps in technology" and other innovation inventories, *Scientometrics*, 1 (1979) 85–98; *Scientific Productivity*, F. M. ANDREWS Ed., Cambridge Univ. Press, and UNESCO, Cambridge, Paris, 1979; *Sociology of Science and Research*, J. FARKAS Ed. Budapest Akadémiai Kiadó, 1979; M. P. CARPENTER, F. NARIN. The subject composition of the world's scientific journals, *Scientometrics* 2 (1980) 53–83; A. HEERINGEN. Dutch research groups, output and collaboration, *Scientometrics*, 3 (1981) 205–315.
20. J. S. COLEMAN, *Introduction to mathematical sociology*, Free Press of Glencoe, N. Y., 1964; *Sotsiologiya v SSSR (Sociology in the USSR)*, v. 1, 2 Moscow, Mysl, 1966; F. EDDING, D. BERSTECHER, *International developments of educational expenditure 1950–1965*, UNESCO, Paris, 1969; B. HILL, op. cit., note 9; M. G. BULMER, On fitting the Poission longormal distribution to species abundance data, *Biometrics*, 30 (1974) 101–110; R.

MORGAN, E. E. IRONS, E. A. PEREZ, T. N. SOULE, A. K. FRIED *Science and Technology for Development*, N. Y., Pergamon Press, 1979.

21. See S. D. HAITUN, op. cit., note 1, part III, p. 190–191.
22. See S. D. HAITUN, op. cit., note 1, part II, p. 94.