INFORMETRICS 89/90
L. Egghe and R. Rousseau (Editors)

CRITERIA OF GAUSSIAN/NON-GAUSSIAN NATURE OF DISTRIBUTIONS AND POPULATIONS

S.D. HAITUN

Institute of the History of Science and Technology
USSR Academy of Sciences
Staropansky per., 1/5, Moscow, 103012, USSR

Abstract

The use of Gaussian distributions in approximations of samples of non-Gaussian populations leads to irreproducible results. Non-Gaussian distributions should be used in these cases. The criteria of Gaussian/non-Gaussian nature for distributions are distinct, thus making it possible to unequivocally show which distribution and at which values of parameters is Gaussian/non-Gaussian. For populations, the criteria are more blurred. An alternative is to use only Zipfian distributions as approximations. The correctness of various distributions (GIGP, GW etc) and the methods of their use in informetrics and scientometrics is analysed.

1. INTRODUCTION

In a number of publications [1-3] we have developed the notions of the non-Gaussian nature of man. The idea is that unlike the natural distributions, the stationary (i.e. containing no time) distributions describing human and, specifically, scientific activities very frequently have characteristic long tails. The distributions obtained using closed scales (an example of which can be school marks) are Gaussian (short-tailed). However, the closed scales deforming the indicator scale with respect to the latent one are incorrect and should be replaces with open scales [4]. That is when the non-Gaussian distributions become dominant in science studies. Bibliometrics, scientometrics and informetrics as a rule use open scales and thus the distributions occuring here are mostly non-Gaussian.

For non-Gaussian distributions the moments increase as the sample size goes up. So far as the sample size randomly changes from researcher to researcher, the moments and whatever is based on them cannot be used for these distributions. Specifically, non-Gaussian data should not be approximated by Gaussian distributions the parameters of which are determined by the moments and, therefore, fluctuate depending on the sample size. Thus, in the case of the Gauss or Poisson distribution the distribution parameters directly coincide with the moments. For a negative binomial distribution

$$n(x) = \frac{\Gamma(x+k)}{\Gamma(x)\,\Gamma(x+1)}\,\theta^x,\ 0 < \theta < 1,\ k > 0,\ x = 1,2,3,... \qquad (1)$$

the mean and the variance can be expressed via distribution parameters k and θ :

$$\bar{x} = \frac{k^\theta}{(1-\theta)\,[1-(1-\theta)^k]}$$

$$\overline{x^2} = \frac{k\theta\ (1+k\theta)}{(1-\theta)^2\ [1-(1-\theta)^k]} + \frac{k}{(1-\theta)\ [1-(1-\theta)^k]} \quad . \qquad (2)$$

Evidently, the inverse is also true : we can express parameters k and θ in terms of $\overline{x}$ and $\overline{x^2}$. And if, as it should be for the non-Gaussian data, the sample moments increase as the sample size goes up, then sample parameters $\hat{k}$ and $\hat{\theta}$ will also change depending on the sample size in one and the same population. And thus the researchers operating on one and the same population and approximating the data by one and the same negative binomial distribution will obtain different values of parameters for their approximations. The situation with geometric, logarithmic, lognormal etc. Gaussian distributions is absolutely identical.

What is meant is the necessity of developing a novel view of empirical data approximation. Up to the present, the research effort has been focussed on an as-successful-as-possible approximation of a *particular sample*. What I suggest is to also recall that approximation parameters should be reproducible, i.e. one should strive to achieve an as-successful-as-possible approximation of *all possible samples* from a given population. Otherwise, we would toil each in his own nook until the end of time obtaining each his own results.

To approximate non-Gaussian empirical data by only non-Gaussian distributions one should know, on the one hand, which data are Gaussian and which are not and, on the other, which distributions are Gaussian and which are not. In other words one should have the criteria of Gaussian/non-Gaussian nature. This paper analyzes such criteria.

2. CRITERIA OF GAUSSIAN/NON-GAUSSIAN NATURE OF DISTRIBUTIONS

Let us give a strict definition of the non-Gaussian distribution. The basis of modern applications of mathematical statistics and the theory of probability are the limit theorems of the convergence of sum distributions of equally distributed random independent values to certain finite so-called stable distributions.

The stable distributions are those the convolution of which with the same distributions results in a distribution of the same kind, i.e. if for any $a_1 > 0$, b_1, $a_2 > 0$ and b_2 such $a > 0$ and b will be found that for all x

$$F(a_1x + b_1) * F(a_2x + b_2) = F(ax + b) \ . \qquad (3)$$

The convolution of the two distributions $f_1(\xi_1)$ and $f_2(\xi_2)$

$$\int f_1(y-x)\ f_2(x)\ dx \qquad (4)$$

is the sum distribution $\eta = \xi_1 + \xi_2$ of the random variables ξ_1 and ξ_2 described by the convoluted distributions.

The distributions pertaining to the stable types we shall divide into two classes : Gaussian and non-Gaussian. Following Yablonsky [5-6] we shall call Gaussian the distributions converging in the above sense to the Gauss distribution; non-Gaussian, those converging to other stable distributions.

The Gaussian distributions, according to the central limit theorem, have the first two moments finite; the non-Gaussian distributions, infinite. And this gives us the first criterion of the Gaussian nature of a distribution : if we succeed in analytically expressing the mean and the variance in a finite form via distribution parameters, the distribution is Gaussian. Just this is the case with the above-mentioned Gauss, Poisson, negative binomial, geometric, logarithmic, lognormal distributions. The same is true for Good distributions [7],

$$f(x) = \{\theta/[\theta + (1-\theta)\ \ln(1-\theta)]\}\theta^x/x(x-1)\ ,$$
$$x = 1,2,3,\ldots,\ 0 < \theta < 1 \qquad (5)$$
$$\overline{x} = \frac{\theta + \ln(1-\theta)}{\theta + (1-\theta)\ \ln(1-\theta)}\ , \qquad \overline{x^2} = 1/(1-\theta)$$

and

$$f(x) = (-\ln\theta)^{s-1}\ \Gamma(1-s,\ -x_0\ \ln\theta)x^{-s}\ \theta^x\ ,$$
$$x \geq x_0,\ 0 < \theta < 1\ , \qquad (6)$$
$$\overline{x} = \frac{1}{(-\ln\theta)}\frac{\Gamma(2-s,\ -x_0\ \ln\theta)}{\Gamma(1-s,\ -x_0\ \ln\theta)}\ ,$$
$$\overline{x^2} = \frac{1}{(-\ln\theta)^2}\frac{\Gamma(3-s,\ -x_0\ \ln\theta)}{\Gamma(1-s,\ -x_0\ \ln\theta)}$$

($\Gamma(\nu,\mu)$ is the incomplete gamma function) and some other distributions.

Sometimes one fails to immediately establish the analytical type of the distribution due to its complexity. Of use then could prove the study of the distribution asymptotics at large values of random variable x, the basis being the Gnedenko-Doeblin limit theorem. The theorem reads that for the given distribution F(x) to converge in the above sense to a stable distribution differing from the Gauss distribution, it is necessary and sufficient that at $x \to \infty$

$$F(-x) \sim C_1\frac{h_1(x)}{|x|^\beta}\ , \qquad 1 - F(x) \sim C_2\frac{h_2(x)}{x^\beta}\ ,$$
$$C_1 > 0,\ C_2 > 0,\ C_1 + C_2 > 0,\quad 0 < \beta < 2 \qquad (7)$$

Here, F(x) is an integral distribution function $F(x) = \int f(x)dx$; and $h_i(x)$ are functions slowly varying in the sense of Caramat, i.e. such that for all $t > 0$

$$\lim_{x\to\infty}\frac{h_i(tx)}{h_i(x)} = 1\ . \qquad (8)$$

In other words, with the accuracy of up to the slowly varying function the Zipf distribution

$$n(x) = \frac{C}{x^{1+\alpha}}\ ,\ x \geq x_0\ ,\quad 0 < \alpha < \infty\ , \qquad (9)$$

at $\alpha < 2$ coincides with the asymptotic of non-Gaussian distributions.

Hence the second criterion of the Gaussian/non-Gaussian nature of distributions : if we succeed in determining the asymptotics type of a distribution f(x)

$$f(x) \underset{x\to\infty}{\sim} \frac{1}{x^{1+\alpha}} , \tag{10}$$

then at $\alpha < 2$ the distribution is non-Gaussian; at $\alpha > 2$, Gaussian.

The distributions which at large values of x have the form of the Zipf distribution we shall call *Zipfian*; however, if in (10) $\alpha = \infty$ then the distribution is *non-Zipfian*. The Zipfian distribution at $\alpha < 2$ is non-Gaussian; at $\alpha > 2$, Gaussian. The Gaussian non-Zipfian distributions have the asymptotics with $\alpha = \infty$. The Gauss, Poisson, negative binomial and other above-mentioned Gaussian distributions are just non-Zipfian. It can be seen that the mean and the variance expressed for them via distribution parameters are finite at all admissible values of these parameters. *If the mean and the variance* of a Zipfian distribution are expressed via distribution parameters including α, then they will prove to be finite only for $\alpha > 2$.

In the log-log coordinates for a Zipfian distribution, Gaussian or non-Gaussian we have :

$$\lim_{x\to\infty} \frac{d\ln f(x)}{d\ln x} = -(1+\alpha) \tag{11}$$

For a Gaussian non-Zipfian distribution :

$$\lim_{x\to\infty} \frac{d\ln f(x)}{d\ln x} = -\infty . \tag{12}$$

Let us adduce examples of using this criterion. Bulmer [8] uses the lognormal *Poisson distribution*

$$f(x) = \frac{(2\pi V)^{-1/2}}{x!} \int_0^\infty \lambda^{x-1} e^{-\lambda} e^{-(\ln x-M)^2/2V} d\lambda , \tag{13}$$

$x = 0,1,2,\ldots$, to describe the species abundance distribution. For this distribution, as it can be shown,

$$\lim_{x\to\infty} \frac{d\ln f(x)}{d\ln x} = -\infty , \tag{14}$$

thus it is Gaussian non-Zipfian.

The generalized inverse Gauss-Poisson (GIGP) or Sichel distribution [9,10] finds the ever growing application in descriptions of social empirical distributions [11]. It has the following form :

$$f(x) = \frac{(1-\theta)^{\nu/2}}{K_\nu\,[\beta(1-\theta)^{1/2}]} \frac{(\frac{\beta\theta}{2})^x}{x!} K_{x+\nu}(\beta) , \tag{15}$$

$$x = 0,1,2,\ldots, \quad 0 \le \theta \le 1 , \quad \beta \ge 0 ,$$

where $K_n(z)$ is a modified Bessel function of the second kind of the order n with argument z. Leaving the first term in the expansion of the Bessel function $K_n(z)$ *at* $n \to \infty$ [12] *we find for its asymptotics* :

$$K_n(z) \underset{n\to\infty}{\sim} (n-1)!(\tfrac{n}{2})^{-n} . \tag{16}$$

For the asymptotics of the Sichel distribution it gives :

$$f(x) \underset{x\to\infty}{\sim} \frac{\theta^x}{x^{1-\nu}}, \quad \lim_{x\to\infty} \frac{d\ln f(x)}{d\ln x} = -(1-\nu) + x\ln\theta , \tag{17}$$

i.e.

$$\lim_{x\to\infty} \frac{d\ln f(x)}{d\ln x} = -\infty \quad \text{at } \theta < 1 \tag{18}$$

and

$$f(x) \underset{x\to\infty}{\sim} \frac{1}{x^{1-\nu}} \quad \text{at } \theta = 1 . \tag{19}$$

Thus, at $\theta = 1$, ν can be only negative (otherwise $f(x)$ would not yield to the normalization); then the Sichel distribution is Zipfian with $\alpha = -\nu$; at $\nu < 2$ it is Gaussian; at $\nu > 2$, non-Gaussian. At $\theta < 1$ it is Gaussian non-Zipfian; at $\beta = 0$, $\nu > 0$ it passes over to the negative binomial distribution; at $\beta = 0$, $\nu = 0$, to a logarithmic distribution [10].

The use of the generalized Waring (GW) distribution [13,14] in bibliometrics is suggested by Burrell [15]. This distribution has the following form :

$$f(x) = \frac{\Gamma(a+c)}{B(a,b)\,\Gamma(c)} \frac{\Gamma(x+c)\,\Gamma(x+b)}{\Gamma(x+a+b+c)} \frac{1}{x!} , \tag{20}$$

$$x = 0,1,2,\ldots .$$

As it can be easily seen, for it

$$f(x) \underset{x\to\infty}{\sim} \frac{1}{x^{1+a}} . \tag{21}$$

i.e. this distribution is Zipfian with $\alpha = a$; at $a < 2$ it is non-Gaussian; at $a > 2$, Gaussian.

Price [16] makes use of the beta function distribution

$$n(x) = N(m+1) \frac{\Gamma(x)\,\Gamma(m+2)}{\Gamma(x+m+2)} . \tag{22}$$

For it

$$n(x) \underset{x\to\infty}{\sim} \frac{1}{x^{2+m}} . \tag{23}$$

Thus, this distribution is Gaussian at $m > 1$ and non-Gaussian at $m < 1$.

To describe the distribution of persons by income, Davis [17] utilizes the generalized Pareto distribution :

$$f(x) \sim \frac{1}{e^{b/x} - 1} \frac{1}{x^{\beta}} . \tag{24}$$

For it

$$f(x) \underset{x\to\infty}{\sim} \frac{1}{x^{\beta-1}} , \tag{25}$$

and the given distribution is Zipfian with $\alpha = \beta - 2$; at $\beta < 4$ it is non-Gaussian; at $\beta > 4$, Gaussian.

The Zipf distribution itself (9) is Gaussian at $\alpha > 2$ and non-Gaussian at $\alpha < 2$.

Until now we spoke about the frequency form of the distribution f(x) when x is understood as a random variable and f(x), the probability density. The rank form of distribution requires special discussion [18]. Here we will confine ourselves to the following.

Any sample distribution can be presented both in the frequency form f(x) or F(x) and the rank form x(r) or X(r). The rank form is introduced by the ratio :

$$r(x) = \sum_{\xi=x}^{J} n(\xi), \quad \sum_{x_0}^{J} n(\xi) = N , \quad 1 \le r \le N . \tag{26}$$

Rank r(x) here is the ordinal number of a given value of random variable x when these values are arranged in the order of decreasing x, all the entities having different ranks. That is why the maximal rank r_{max} equals to the sample size N.

The rank differential form x(r) is determined directly (25). The rank integral form,

$$X(r) = \sum_{1}^{r} x(\xi) , \quad \sum_{1}^{N} x(\xi) = G . \tag{27}$$

The Mandelbrot distribution [19]

$$x(r) = \frac{A}{(r+B)^{\gamma}} , \tag{28}$$

as it can be easily shown, is a rank form of the Zipf distribution (9) with $\alpha = 1/\gamma$. It is Gaussian at $\gamma < 1/2$ and non-Gaussian at $\gamma > 1/2$.

The Bradford distribution [20]

$$X(r) = a + b\log r , \tag{29}$$

the Cole distribution [21],

$$X(r) = X(N)\ (1 + a \log \frac{r}{N}) , \tag{30}$$

and the Leimkuhler distribution [22]

$$X(r) = \frac{a\log (1+br)}{\log (1+b)} \tag{31}$$

are also Zipf distributions with the fixed value of parameter $\alpha = 1$. Therefore, these three distributions are non-Gaussian.

3. CRITERIA OF THE GAUSSIAN/NON-GAUSSIAN NATURE OF POPULATIONS

The 'infinite' moments are an abstraction suitable only for theoretical distributions mentally plotted on infinite size populations. In reality one should consider the 'essential' or 'non-essential' growth of the moments with respect to the population sample size. If this growth is essential the population is non-Gaussian; non-essential, Gaussian.

Hence, the first criterion of the Gaussian/non-Gaussian nature of a population. One has 'merely' to plot the random samples of different sizes N from a given population and check the dependence of the mean and the variance on N. If it is there, the population is non-Gaussian; if it is negligible, Gaussian.

However, this procedure is complicated and one has to resort to simplified criteria.

In the analysis of the data by Sichel [9] we made use of the following technique. His paper includes the table data on nine empirical distributions of sentence length in the texts by Herodotes, Macauley, Wells etc. Those distributions have approximately the same form an idea of which is given in figure 1. We approximated them graphically at large values of x by the Zipf distribution and for each of them determined the value of index α by the slope of the asymptote in the log-log coordinates to the abscissa axis. The total data on nine distributions are given in table 1. The dependence of σ^2 vs $\bar{x}$ for nine samples considered was plotted (figure 2). The plot shows that two samples clearly drop out of the pattern and thus could not be considered as the samples of one and the same population. For the other samples the dependences $\bar{x}$ and σ^2 on N are shown in figure 3. The plot indicates that these seven samples are disintegrated according to the values of α into three groups, the first of which contains only one sample. We can see that in the case of the second and third groups there exists the pronounced dependence of $\bar{x}$ and σ^2 on N which proves the non-Gaussian nature of this population.

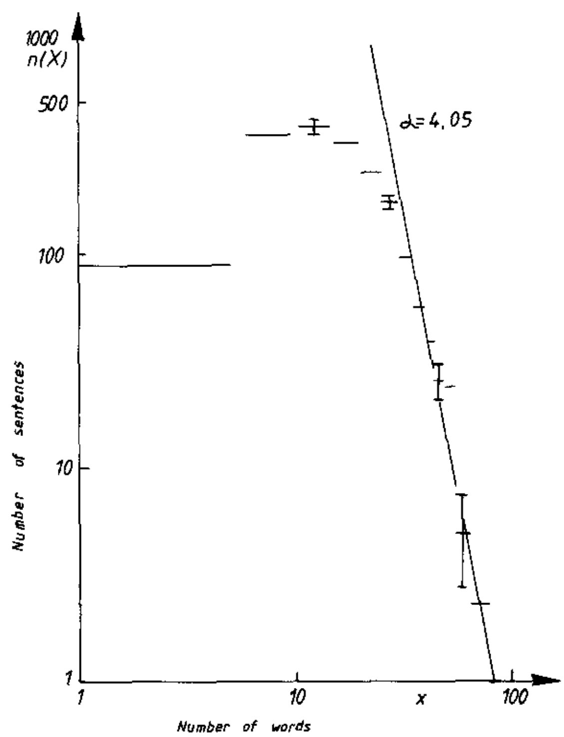

Fig.1 : Sentence-length distribution in the excerpts from the texts by Herodotes (the data from Morton). The sampling is 1800 sentences. Log-log coordinates, the straight line corresponds to the Zipf distribution. The value of parameter α of this distribution, by which the data at high values of the variable were approximated, is determined graphically (cf.fig.5).

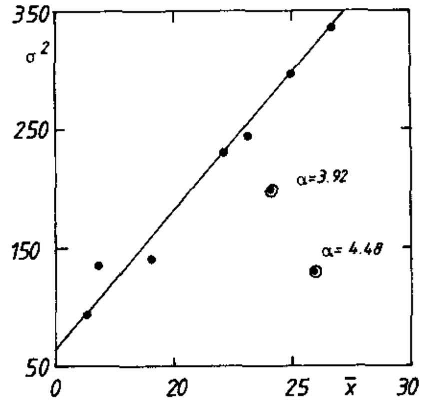

Fig.2 : Dependence of $\sigma^2(\bar{x})$ for nine Sichel samples

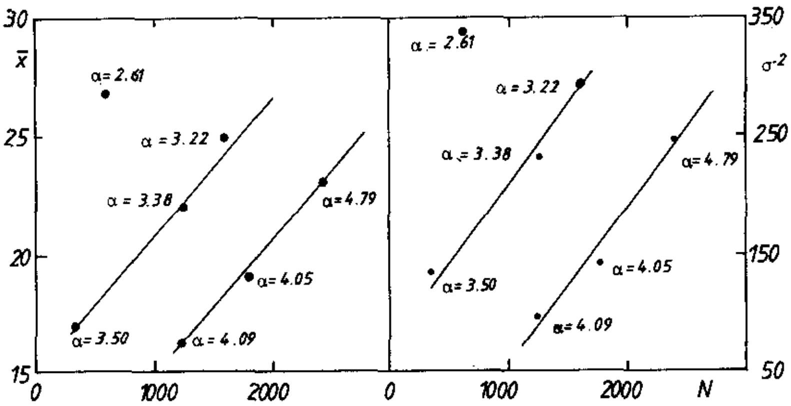

Fig.3 : Dependence of the mean $\bar{x}$ and the variance σ^2 on the sample size N for nine Sichel samples. The plots indicate an essential dependence of $\bar{x}$ and σ^2 on N for the data by Sichel and, therefore, the non-Gaussian nature of the non-Gaussian population studied by him.

Table 1

N	355	600	600	625	1221	1251	1600	1800	2417
$\bar{x}$	16.96	24.09	25.91	26.77	16.24	22.07	24.98	19.04	23.07
σ^2	133.56	199.38	131.05	337.24	96.36	230.72	292.73	140.85	244.98
α	3.50	3.92	4.48	2.61	4.09	3.38	3.22	4.05	4.79

The second criterion. Rather frequently scientometric and, generally, social distributions appear to yield to, if they are based on small size samples, the satisfactory approximation by the Gaussian non-Zipfian distribution. Often the lognormal distribution

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)^2\right], \quad x \geq 0 \tag{32}$$

plays this role due to its relative long-tailedness; empirical points in this case lay well on the line corresponding to this distribution in the log probability coordinates. The conclusions of the non-Gaussian nature of the given population can be arrived at in this case by two ways. First, one can increase the sample size and then, if the population is indeed non-Gaussian, the tail of the empirical distribution sooner or later would deviate from the straight line. Second, one can construct for this theoretical (say, lognormal) distribution the plots of the dependence of the mean $\overline{x}$ and the variance σ^2 on the maximal sample value $x_{max} = J$,

$$\overline{x}(J) = \frac{\sum_{x_0}^{J} xn(x)}{\sum_{x_0}^{J} n(x)}, \quad \sigma^2 = \frac{\sum_{x_0}^{J} (x-\overline{x})^2 n(x)}{\sum_{x_0}^{J} n(x)}. \tag{33}$$

If the dependences $\overline{x}(J)$ and $\sigma^2(J)$ show the essential growth near the sample value $J = \hat{J}$, we should consider our population non-Gaussian; if near $J = \hat{J}$ they come to a plateau, the population is Gaussian. Here we proceed from the fact well-known in mathematical statistics that during the increase in the sample size N at the fixed distribution f(x) occurring on the population the growth of J is most probable as well, so that the dependence of $\overline{x}$ and σ^2 on J also indicates that on N. An example of the use of this criterion is shown in figure 4.

The third criterion. We approximate the data at large values of x by the Zipf distribution and by the slope of the asymptote in the log-log coordinates to the abscissa axis we determine the Zipf distribution index α (figure 5). The value of α we determine more often graphically because the situation with the use of statistical methods of determining the distribution parameters in a non-Gaussian case is far from being clear. If α is small, we consider the distribution non-Gaussian; large, Gaussian. What value of α could be considered as 'small' and what is 'large', is impossible to say unequivocally. Practically we believe a population to be non-Gaussian if $\alpha < 2$; Gaussian, if $\alpha \gtrsim 6$. In the intermediate region one should resort to extra measures. Say, to the second criterion.

On the whole it should be admitted that the situation with the criteria of the Gaussian/non-Gaussian nature of populations is considerably more complicated than with those of distributions. The strict solution of the problem requires in each particular case a rather labour-consuming analysis. The analysis which, besides all other things, should also take into account the degree of reproducibility of results sufficient in each concrete investigation; and that by itself brings in an additional indefiniteness.

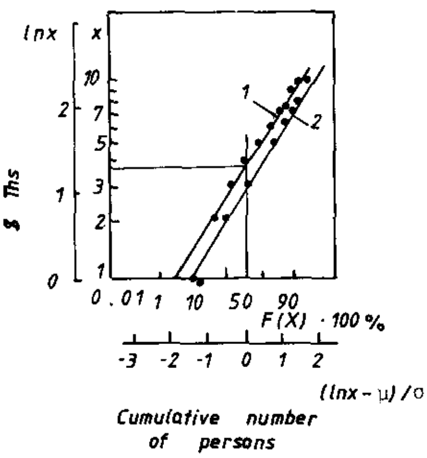

Fig.4a : Distribution of the USA population in 1944 and 1950 by private income [23]. 1,1950; 2,1944.Log probability coordinates, the straight line corresponds to the lognormal distribution. Parameters of one of them (1950) : $\sigma = 0.66$; $\mu = -1.94$. The lognormal approximations here are, however, incorrect. First, this plot shows no distribution tail corresponding to large personal income and deviating from the lognormal straight line. Second, in the area of x values where the lognormal straight line occurs, this approximation is incorrect anyway (cf. fig.4b).

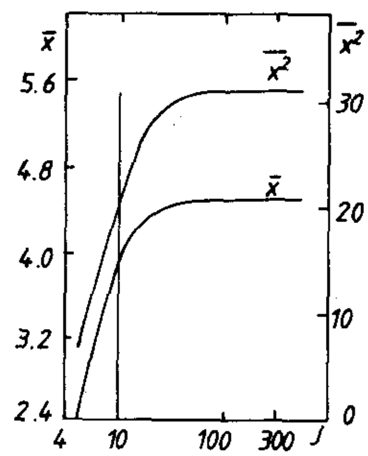

Fig.4b : Dependences of the mean $\overline{x}$ and $\overline{x^2}$ on the maximal sample value of the variable $x_{max} = J$ for the lognormal distribution with the values of parameters σ and μ determined graphically by figure 4a. As we can see, in the area of the sample value $\hat{J} = 10$ these dependences are essential which makes the lognormal approximation incorrect.

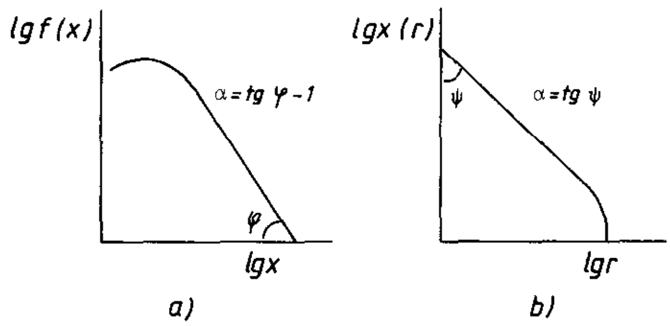

Fig.5 : Determination of parameter α of the Zipfian distribution by the slope of the asymptote to the coordinate axes at $x \to \infty$ in the log-log coordinates

4. SELECTION OF APPROXIMATIONS

Non-Gaussian populations, we said in the first section, should be approximated by non-Gaussian distributions otherwise we would not be able to ensure the reproducibility of the results necessary for the investigation. The criteria of the Gaussian/non-Gaussian nature of distributions, as we saw in the second section, have the distinct character and their practical use causes no difficulties. Those of populations, however, are on the opposite extremely labour-consuming and, in fact, blurred. At the same time, if one uses only open scales practically all empirical distributions occurring in social sciences are Zipfian with varying values of α. All the more so in scientometrics and informetrics. That is why *we suggest to use here only the Zipfian distributions* and not apply at all the criteria of the Gaussian/non-Gaussian nature of populations. If we will make use of only the Zipfian distributions and for the treatment of the results will utilize only those statistical methods which are not based on the moments (i.e. the methods of non-Gaussian mathematical statistics) we will automatically avoid the dependence of the results on the sample size. By this we will eliminate the source of irreproducibility of the results due to the non-Gaussian nature of scientific activities.

All Zipfian distributions contain parameter α or its equivalent. The problem is, thus, the choice of a Zipfian approximation which apparently would more successfully describe the area of the small values of variable x. The beta function distribution contains no parameter responsible for this area and thus is maximally unsuitable. The Pareto generalized distribution contains one, which makes it more suitable. The Waring generalized distribution includes two such parameters and this, in fact, ensures its highest flexibility though increases the complexity of calculations.

As for the Sichel distribution, certain doubts can be expressed in its respect. At present, the value of parameter ν of this distribution, following Sichel, is postulated as a rule to be equal to -1/2 and the sample values of parameters β and θ are determined by the sample mean and variance using the expressions [9]

$$\bar{x} = \frac{\beta\theta}{2(1-\theta)^{1/2}}, \qquad \sigma^2 = \frac{\beta\theta(2-\theta)}{4(1-\theta)^{3/2}} \tag{34}$$

which in most cases gives $\theta < 1$.

In our view, not everything is correct here. When passing over from one empirical distribution to another α, generally speaking, changes and thus ν ($= -\alpha$) cannot in principle be postulated here as equal to $-1/2$ for all cases. Moreover, no value of ν could be postulated here at all. In each particular case its concrete value should be determined. However, if even by chance the value of $\nu = -1/2$ would have proven suitable for some empirical distribution, the calculation of the values of parameters β and θ via $\bar{x}$ and σ^2 is incorrect because at $\alpha = 1/2$ the distribution is essentially non-Gaussian, its moments rapidly grow as the sample size increases and thus the calculated values of parameters β and θ prove to be irreproducible. Finally, at $\theta < 1$ the Sichel distribution, as we saw, is Gaussian non-Zipfian so it would be altogether inexpedient to use it for approximation of bibliometric distributions having such θ.

The correct use of the Sichel distribution could consist in the following. First of all, postulate $\theta = 1$, i.e. make the Sichel distribution Zipfian with $\alpha = -\nu$. Further, using the statistical methods of determining the distribution parameters, select the values of ν characterizing the tail part and β fixing the form of the distribution in the area of small values of x. There exist specific problems here which we would not dwell on now. We would only point out that, say, the maximal likelihood method and, apparently, chi-squared *method significantly overestimate* α. Anyway, the Bessel function index which includes parameter ν is not a very suitable object to vary using any procedure of determining the distribution parameters. Due to this reason the choice of the Sichel distribution appears to be generally not the most appropriate for use in bibliometric and scientometric applications.

REFERENCES

[1] Haitun, S.D., The stationary scientometric distributions. Part 1 : Different approximations. *Scientometrics*, 4 : 5-25, 1982. Part 2 : Non-Gaussian nature of scientific activities, 4 : 89-104, 1982. Part 3 : Role of the Zipf distribution. Scientometrics, 4 : 181-194, 1982.

[2] Haitun, S.D., Naukometriya : Sostoyanie i perspektivy (Scientometrics : State and Prospects). Moscow, Nauka, 1983, 344 p.

[3] Haitun, S.D., On M. Kunz's article "A case study against Haitun's conjectures". Scientometrics, 13 : 35-44, 1988.

[4] Haitun, S.D., Problems of quantitative analysis of scientific activites : *Non-additivity of data*. Part 1 : Statement and solution. Scientometrics, 10 : 3-16, 1986. Part 2 : Corollaries. Scientometrics, 10 : 133-135, 1986.

[5] Yablonsky, A.I., On fundamental regularities of the distribution of scientific productivity. Scientometrics, 2 : 3-34, 1980.

[6] Yablonsky, A.I., Stable non-Gaussian distributions in scientometrics. Scientometrics, 7 : 459-470, 1985.

[7] Good, I.J., The population frequencies of species and the estimation of population parameters. Biometrics, 40 : 237-264, 1953.

[8] Bulmer, M.C., On fitting the Poisson lognormal distribution to species - abundance data. Biometrics, 30 : 101-110, 1974.

[9] Sichel, H.S., On a distribution representing sentence -length in written prose. Journal of the Royal Statistical Society, A137 : 25-34, 1974.

[10] Sichel, H.S., A bibliometric distribution which really works. Journal of the American Society for Information Science, 36 : 314-321, 1985.

[11] Stein, G.Z., Zucchini, W., Juritz, J.M., Parameter estimation for the Sichel distribution and its multivariate extension. Journal of the American Statistical Association, 82 : 938-944, 1987.

[12] Dwight, H.B., Tables of integrals and other mathematical data. New York : Macmillan, 1961.

[13] Irwin, J.O., The generalized Waring distribution applied to accident data. Journal of the Royal Statistical Society, A131 : 205-225, 1968.

[14] Irwin, J.O., The generalized Waring distribution. Journal of the Royal Statistical Society, A138 : 18-31, 1975 (Part 1); 204-227 (Part 2); 374-384 (Part 3).

[15] Burrell, Q.L., Modelling the Bradford phenomenon. Journal of Documentation, 44 : 1-18, 1988.

[16] Price, D. De Solla, A general theory of bibliometric and other cumulative advantage processes. Journal of the American Society for Information Science, 27 : 292-306, 1976.

[17] Davis, H.T., The analysis of economic time series. Bloomington (Ind.) : Princeton Press, 1941.

[18] Haitun, S.D., The "rank-distortion" effect and non-Gaussian nature of scientific activities. Scientometrics, 5 : 375-395, 1983.

[19] Mandelbrot, B., Contribution a la théorie mathématique des jeux de communication. Philos. D. Thesis, Paris, dec.16, 1952, Publs. Inst. Statist. Univ. Paris, 2 (1/2) : 80-102, 1953.

[20] Bradford, S.C., Documentation. London : Grosby Lockwood and Son Ltd, 1948.

[21] Cole, P.F., A new look at references scattering. Journal of Documentation, 18 : 58-64, 1962.

[22] Leimkuhler, F.F., The Bradford distribution. Journal of Documentation, 23 : 197-207, 1967.

[23] Aitchison, J., Brown, J.A., The lognormal distribution. Cambridge : Univ. Press, 1957.