Chapter 4
Frequency and Rank Approaches to Research Production. Classical Statistical Laws

Dedicated to Lotka, Zipf, Pareto, Mandelbrot, Bradford, Price, and all others who contributed to the study of non-Gaussian effects in the research area of scientometrics

Abstract We discuss several classical statistical laws that are important for understanding characteristics of research production and for its assessment. The statistical laws are grouped in such a way that the two much-used statistical approaches for the study of research systems and especially for the study of research publications (frequency approach and rank approach) are appropriately addressed. We begin with some remarks on the frequency and rank approaches to distributions and discuss why the frequency approach is much used in the natural sciences and the rank approach is widely used in the social sciences. Then the stable non-Gaussian distributions are described, and their importance for statistical methodology of research dynamics is emphasized. The laws of Lotka, Pareto, Zipf, Zipf–Mandelbrot, and Bradford are discussed from the point of view of their application to describing different aspects of scientific production. In addition to the discussion of statistical laws, we discuss two important effects: the concentration–dispersion effect (which reflects the separation of the researchers into a small group of highly productive ones and a large group of researchers with limited productivity) and the Matthew effect in science (which reflects the larger attention to the research production of the highly ranked researchers). In addition, we mention the invitation paradox (many papers accepted in highly ranked journals are not cited as much as expected) and the Ortega hypothesis (the big discoveries in science are supported by the everyday hard work of ordinary researchers). At the end of the chapter we discuss more general questions; relationships between the statistical laws and power laws as informetric distributions.

4.1   Introductory Remarks

The action of various "soft" laws may be observed in the area of research dynamics. An example of such a law is the principle of cumulative advantage formulated by Price [1]: *Success seems to breed success. A paper which has been cited many times is more likely to be cited again than one which has been little cited. An author of many papers is more likely to publish again than one who has been less prolific. A journal which has been frequently consulted for some purpose is more likely to be turned to again than one of previously infrequent use.* Our attention is concentrated in this book on research publications as units of scientific information and on citations of research publications as units for impact of the corresponding scientific information. Below, we discuss several statistical power laws connected to research publications and their citations. We emphasize the fact that the discussed power laws should be considered statistical laws ("soft" laws), i.e., more as trends and not as laws that are similar to the "hard" laws of physics. Because of this, one could expect that deviations from the discussed power laws will occur in some real situations. There is a large amount of literature devoted to application of different power laws for modeling features of research dynamics [2–11], and this literature is a part of the literature devoted to the applications of power laws in different areas of science [12–16]. From the point of view of mathematics, the statistical laws connected to research publications and citations are very interesting, since these laws are described mathematically by the same kinds of relationships (hyperbolic relationships),[1] which is evidence of a general structural mechanism of research organizations and scientific systems [17].

The regularities discussed below describe a wide range of phenomena both within and outside of the information sciences. These regularities (called laws and named after the prominent researchers associated with them) were observed in many research fields in the last century. Below, we shall discuss mainly regularities connected to research publications. Let us note that the discussed statistical laws occur in many other areas, such as linguistics, business, etc (Figs. 4.1 and 4.2).

4.2   Publications and Assessment of Research

> *The pure and simple truth is rarely pure and never simple*
> Mark Twain

Research production is evaluated often by indicators and indexes connected to research publications [18–21]. There are interesting relationships connected to publications and their authors. These relationships are based on the existence of regularities in the publication activity of the authors of publications. The first relationship

---

[1]Hyperbolic relationships are relationships of type $m_i i^\alpha = $ const. Such relationships are frequently observed in different areas of science such as biology and physics. They exist also in the area of mathematical modeling of structures, processes, and systems in the area of social sciences.
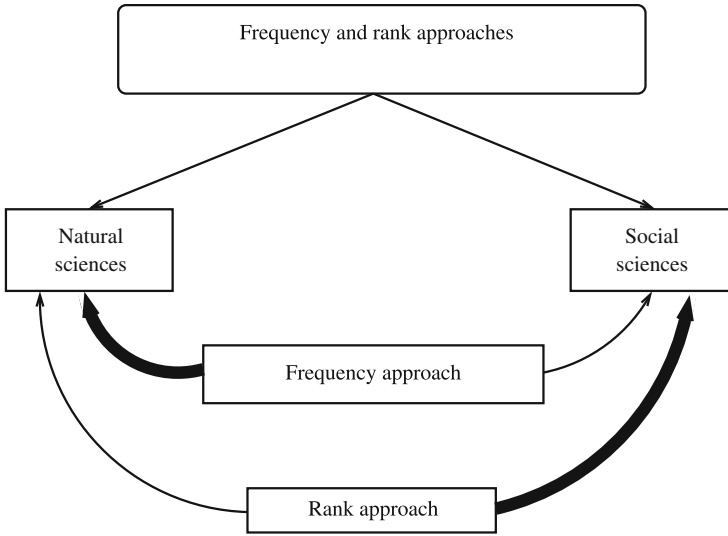
Fig. 4.1 The frequency approach is dominant in the natural sciences. The rank approach is much used in the social sciences
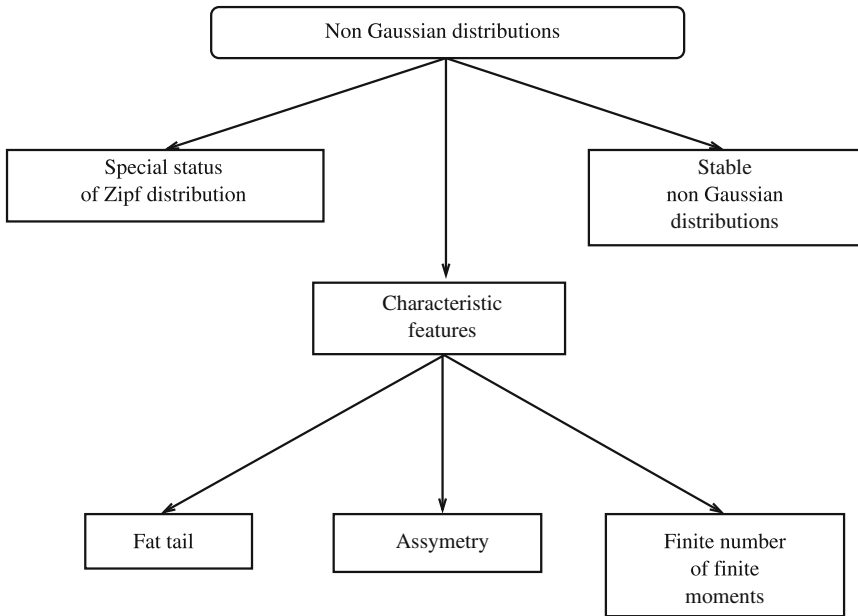

Fig. 4.2 The Zipf distribution has a special status in the world of non-Gaussian distributions (and this status is close to the status of the normal distribution in the world of Gaussian distributions). Non-Gaussian distributions have interesting features that have even more interesting consequences. Stable non-Gaussian distributions arise frequently in different areas of science

was discovered in 1926, when Alfred Lotka (the same Lotka known for the famous Lotka–Volterra equations in population dynamics) published an article [22] on the frequency distribution of scientific productivity determined from an index of *Chemical Abstracts*. The conclusion was that *the number of authors making n contributions is about $1/n^2$ of those making one contribution; and the proportion of all contributors who make a single contribution is about* 60 %.

Further discoveries of such kinds of relationships followed. In 1934, Bradford [23] published a study of the frequency distribution of papers over journals. Bradford's conclusion was that *if scientific journals are arranged in order of decreasing productivity on a given subject, they may be divided into a nucleus of journals more particularly devoted to the subject and several groups or zones containing the same number of articles as the nucleus when the numbers of periodicals in the nucleus and the succeeding zones will be as* $1 : b : b^2 : \ldots$. In 1949, Zipf [24] discovered a law in quantitative linguistics (with applications in bibliometrics). This law states that $rf = C$, where r is the rank of a word, f is the frequency of occurrence of the word, and C is a constant that depends on the analyzed text. As we shall see below, this relationship is connected to the relationships obtained by Lotka and Bradford. Zipf also formulated as interesting principle (of least effort) that serves to explain the above relationship: *a person …will strive to solve his problems in such a way as to minimize the total work that he must expend in solving both his immediate problems and his probable future problems…* [24]. In 1963, Price [25] formulated the famous square root law: *Half of the scientific papers are contributed by the top square root of the total number of scientific authors.*

Characteristics of research publications such as their number, type, and distribution are the most commonly applied indicators of scientific output, e.g., the production of a research group is measured often by its number of publications, and productivity is expressed often as the number of publications per person–year [26]. Researchers from different fields of science put different weights on different kinds of publications. Researchers from the natural sciences prefer to publish papers in refereed international journals with (possibly larger) impact factors. Researchers from the humanities prefer to publish results in book form rather than as articles. And researchers from the applied sciences publish their results very often as engineering research reports and patents.

Even within each of the above large fields of science, the weights of the different sorts of the dominant kinds of publications vary. Let us concentrate on the natural sciences and on publications in the form of articles. For a long time, articles have been classified as follows:

1. articles published in journals with impact factor (assigned by SCI (Science Citation Index)) [27–35]. The SCI journals are much cited, highly visible journals for which citation data are available;
2. articles published in journals without impact factor (non-SCI journals). Since the visibility of these journals is smaller compared to the visibility of the SCI journals, publication in non-SCI journals is unlikely to produce the same level of citation.

Because of the above facts, most researchers from the natural sciences have preferred to publish in SCI journals, since publication in such a journal is perceived as a mark of quality of the scientific research. Of interest is that this perception doesn't account for the citation levels, and an uncited article may also be considered a consequence of research of good quality.

Two statistical approaches are much used in the study of sets of research publications and citations: the frequency approach and the rank approach. Let us discuss some of their characteristic features.

4.3 Frequency Approach and Rank Approach: General Remarks

The frequency approach is based on analysis of the frequency of observation of values of a random variable. In the case of research publications, the frequency of observation of a value is the probability that a researcher has written x papers, and the random variable is the production of a researcher from the observed large group of researchers. Such an approach will lead us to the laws of Lotka and Pareto.

The rank approach is based on a preliminary ordering (ranking) of the subgroups (having the same value of the studied quantity) with respect to decreasing values of some quantity of interest. Then one can study the subgroups with respect to their rank. In our case, one can rank the researchers from a large group after building subgroups of researchers having the same number of publications. Such an approach will lead us to the laws of Zipf and Zipf–Mandelbrot. And when we rank the sources of information such as scientific journals, the rank approach will lead us to the law of Bradford. Let us stress here that *a general feature of the laws of Lotka, Pareto, Zipf, and Zipf–Mandelbrot is that these laws are expressed mathematically by hyperbolic relationships.*

The frequency approach and rank approach are appropriate for describing different regions of the distribution of research productivity. The rank approach (the law of Zipf, for example) is appropriate for describing the productivity of highly productive researchers, for which two researchers with the same number of papers rarely exist and the ranking can be constructed effectively. The frequency approach (the law of Lotka, for example) is appropriate for describing the productivity of not so highly productive researchers. This group may contain many researchers, and many of them may have the same number of publications. Because of this, they cannot be effectively ranked, but they can be investigated by statistical methods based on frequency of occurrence of different events (such as number of publications or number of citations).

If the maximum production (the number of publications, number of cita-tions, etc.) of a member of a group of researchers is larger than the number of the members of the group, we may usually use the rank approach for characterization of the research production of these researchers. If the maximum production is much smaller than the number of the members of the group, we have to use the frequency approach.

The frequency and rank statistical distributions have differential and integral forms. Let us consider a large enough sample of items of interest for our study. Let the sample size be N. Let the values of the measured characteristics in the sample vary from x_{min} to x_{max}, and we separate this interval of M subintervals of size $\Delta = (x_{max} - x_{min})/(M)$. Then the differential form of the frequency distribution of x, denoted by $n(x)$ (where n is the frequency of values of x in the interval that contains x), satisfies the relationship

$$\sum_{x_{min}}^{x_{max}} n(x) = N.$$
(4.1)

The integral form of the frequency distribution is

$$f(x) = \frac{1}{N} \sum_{x_{min}}^{x} n(x^*).$$
(4.2)

The differential form of the rank distribution is

$$r = \sum_{x}^{x_{max}} n(x^*), \quad 1 \le r \le N,$$
(4.3)

and the integral form of the rank distribution is

$$R(r) = \sum_{1}^{r} x.$$
(4.4)

Above, the rank means the number of the position of the value x of the studied random variable when all values of the random variable are listed ordered by decreasing frequency $n(x)$.

Let us stress again that in the natural sciences, most of the probability distributions used are frequency distributions. In the social sciences, many of the probability distributions used are rank distributions. But why are frequency distributions dominant in the natural sciences and rank distributions frequently used in the social sciences?

The choice of type of distribution convenient for the statistical description of some sample depends on two factors [36]: the sample size and the value of x_{max}. The

frequency form of the probability distribution is convenient when the normalized frequency $n(x)/N$ is a good approximation of the probability density function. This happens when the frequencies $n(x)$ are large enough and

$$\frac{x_{max} - x_{min}}{\Delta} = M \ll N. \tag{4.5}$$

The corresponding condition for the application of the rank distribution is [36]

$$\frac{x_{min} + x_{max}}{\Delta} \gg 2, \tag{4.6}$$

which means that the rank distributions are more applicable when $\frac{x_{max}}{\Delta}$ is large.

For the case of data from the natural sciences, we usually have large values of N such that the condition (4.5) is satisfied much better that the condition (4.6). In addition, the value of x_{max} is usually not very large. Thus the frequency distributions are dominant. In the social sciences, N is usually not very small, and since the non-Gaussian distributions occur frequently the value of x_{max} is usually large. Thus the condition (4.6) is better satisfied than the condition (4.5), and the rank distributions are used much more than the frequency distributions.

4.4 The Status of the Zipf Distribution in the World of Non-Gaussian Distributions

There is a quotation that if a question is formulated appropriately, that is already half the answer. So let us formulate the question: Why is the status of the Zipf distribution in the world of non-Gaussian distributions almost the same as the status of The Normal distribution is just one distribution from the class of Gaussian distributions?

As we already know, because of the central limit theorem, the normal distribution plays a central role in the world of Gaussian distributions, which are the dominant distributions in the natural sciences. And we know that the non-Gaussian distributions occur frequently in the social sciences. Is there a non-Gaussian distribution that plays almost the same central role for non-Gaussian distributions? There is indeed such a distribution, and its name is the Zipf distribution.

The special status of the Zipf distribution is regulated by the Gnedenko–Doeblin theorem. This theorem [37–40] states that necessary and sufficient conditions (as $x \to \infty$) for convergence of normalized sums of identically distributed independent random variables to stable distributions different from the Gaussian distribution are

$$f(-x) \propto C_1 \frac{h_1(x)}{|x|^{\alpha^*}}; \quad 1 - f(x) \propto C_2 \frac{h_2(x)}{x^{\alpha^*}};$$
$$C_1 \geq 0; \quad C_2 \geq 0; \quad C_1 + C_2 > 0; \quad 0 < \alpha^* < 2, \tag{4.7}$$

where $f(x)$ is the integral frequency form of the corresponding distribution, C_1, C_2, and α^* are parameters, and h_1 and h_2 are slowly varying functions i.e., for all times $t > 0$,

$$\lim_{x \to \infty} \frac{h_k(tx)}{h_k(x)} = 1, \quad k = 1, 2. \tag{4.8}$$

In other words, the Gnedenko–Doeblin theorem states that the asymptotic forms of the non-Gaussian distributions converge to the Zipf distribution (up to a slowly varying function of x).

Let us stress the following.

1. Note the words "up to a slowly varying function." This means that some statistical distributions connected to research publications and citations may deviate from a power law relationship.
2. Note that in the Gnedenko–Doeblin theorem, $\alpha^* < 2$, and for $\alpha^* < 2$, the Zipf distribution is a non-Gaussian distribution. For $\alpha^* > 2$, the Zipf distribution is a Gaussian distribution.
3. When the sample sizes are infinite, the Gaussian distributions have finite moments, and many of the moments of the non-Gaussian distributions are infinite.
4. In practice, one works with finite samples. Then the moments of the Gaussian distributions and the moments of the non-Gaussian distributions may depend on the sample size.

In addition, we note that the statement of the Gnedenko–Doeblin theorem is about the asymptotic form of a non-Gaussian distribution. This has some consequences for the laws (of Lotka, Bradford, etc.) that we shall discuss below. These laws may be considered statistical relationships that are valid for larger sets. In other words, and in most cases (when the studied sets are not large enough), the laws discussed below should be considered trends and not strict rules. These laws are not like the exact 'hard' laws of the natural sciences. However, these laws are stricter than the 'soft' laws that can be found in many of the social sciences.

4.5   Stable Non-Gaussian Distributions and the Organization of Science

Let us recall some characteristic features of non-Gaussian distributions:

(1) **Their "heavy tail"** [41, 42]: This means, for example, that in a research organization there may exist a larger number of highly productive researchers than the normal distribution would lead one to expect.
(2) **Their asymmetry**: There exist many low-productive researchers and not so many high-productive researchers. We shall discuss below that another manifestation of this asymmetry is the concentration–dispersion effect: there is a concentration of productivity and publications at the right-hand side of the Zipf–Pareto distribution, and dispersion of scientific publications among many low-productive researchers at the left-hand side of the distribution.
(3) **They have only a finite number of finite moments**. For example, for the Zipf–Pareto law (with characteristic exponent α), there exist moments of order $n < \alpha$. And if $\alpha = 1$ (as in the case of many practical applications such as the law of Lotka), then there is no finite dispersion.

The nonexistence of the finite second moment violates an important requirement of the central limit theorem (namely the existence of a finite second moment), and thus some distributions do not converge to the normal distribution. Then there is a class of non-Gaussian distributions that describe another "nonnormal" world. And many social and economic systems belong to this world.

The infinite second (and often the infinite first) moment of non-Gaussian distributions means that the probability of large deviations increases, and if the first moment is infinite, then there is no concentration around some mean value.

An important class of non-Gaussian distributions is the class of stable non-Gaussian distributions. The definition of a stable distribution is [43, 44] this: *Suppose that $S_k = X_1 + \cdots X_k$ denotes the sum of k independent random variables, each with the same nondegenerative distribution P. The distribution P is said to be stable if the distribution of S_k is of the same type for every positive integer k. A random variable is called stable if its distribution has this property.*

The normal distribution is a stable distribution. Another class of stable distributions is the class of non-Gaussian distributions with infinite dispersion. And the asymptotic behavior (at $x \to \infty$) of all of these stable non-Gaussian distributions is $\propto \frac{1}{x^{1+\alpha}}$, i.e., convergence to the Zipf–Pareto law.

The origin of the Zipf–Pareto law as the limit distribution for the class of stable non-Gaussian distributions shows that the Zipf–Pareto law reflects fundamental aspects of the structure and operation of many complex organizations om biology, economics, society, etc.

Three stable distributions are known explicitly:

1. The distribution of Gauss (not of interest for us here).
2. The distribution

$$p(x) = \frac{1}{(2\pi)^{1/2}} x^{-3/2} \exp(-\frac{x}{2}),$$ (4.9)

which is connected to a large number of branching processes. At large x, the asymptotic behavior of this distribution is $p(x) \propto \frac{a}{x^{3/2}}$, where $a = (2\pi)^{-1/2}$.

3. The Cauchy distribution [45, 46] (known also as the Lorenz distribution or Breit–Wigner distribution):

$$p(x, x_0, \gamma) = \frac{1}{\gamma \pi \left[1 + \left(\frac{x-x_0}{\gamma}, \right)^2 \right]}$$ (4.10)

where

- x_0: location parameter that specifies the position of the peak of the distribution;
- γ: scale parameter that specifies the half-width at the half-maximum.

Here we shall consider the standard Cauchy distribution $p(x, 0, 1)$, i.e.,

$$p(x) = \frac{1}{\pi} \frac{1}{1 + x^2},$$ (4.11)

whose asymptotic form for large x is $p(x) \propto \frac{a}{x^2}$, where $a = 1/\pi$. We note that for this asymptotic form, we have $\alpha^* = \alpha + 1 = 2$, i.e., $\alpha = 1$. Thus the value of the exponent is the same as the value of the exponent for the law of Lotka for authors and their publications (see the next section). *In other words, the law of Lotka emerges as the asymptotic form of the standard Cauchy distribution.*

4.6   How to Recognize the Gaussian or Non-Gaussian Nature of Distributions and Populations

Usually for non-Gaussian distributions, the moments increase as the the sample size goes up [47]. According to the central limit theorem, the first two moments of Gaussian distributions are finite (which is not the case for the non-Gaussian distri-

butions). Thus the first criterion that a distribution may be Gaussian is *that we are able to express analytically the mean and the variance of the distribution in finite form via distribution parameters*. This is the case of the distributions of Gauss and Poisson, the lognormal distribution, logarithmic distribution, geometric distribution, negative binomial distribution, etc.

The second criterion is connected to the Gnedenko–Doeblin theorem discussed above. The criterion reads thus: *If we are able to determine the asymptotic type of a distribution $f(x)$ and these asymptotics $(x \to \infty)$ are*

$$f(x) \sim \frac{1}{x^{1+\alpha}}, \tag{4.12}$$

then for $\alpha < 2$, the distribution is non-Gaussian, and for $\alpha > 2$, the distribution is Gaussian.

The distributions that at large values of x have the form of a Zipf distribution may be called *Zipfian distributions*. If in (4.12) we have $\alpha = \infty$, then the corresponding distribution is non-Zipfian. The above-mentioned Gaussian distributions are all non-Zipfian distributions. From (4.12), one obtains

$$\lim_{x \to \infty} \frac{d}{d(\ln x)} f(x) = -(1 + \alpha). \tag{4.13}$$

For the Gaussian non-Zipfian distributions, $\alpha = -\infty$.

Two distributions that will be much discussed in the next chapter are the (generalized) Waring distribution and the GIGP (generalized inverse Gauss–Poisson) distribution. It will be useful to know the values of the corresponding parameters for which these distributions are non-Gaussian and/or Zipfian. The GIGP distribution (called also Sichel distribution) is

$$f(x) = \frac{(1 - \theta)^{\nu/2}}{K_\nu[\beta(1 - \theta)^{1/2}]} \frac{(\beta\theta/2)x}{x!} K_{x+\nu}[\beta], \tag{4.14}$$

where $K_n[z]$ is the modified Bessel function of the second kind of order n and with argument z. The asymptotics of $f(x)$ when $x \to \infty$ are given by [47]

$$f(x) \sim \frac{\theta^x}{x^{1-\nu}}. \tag{4.15}$$

Then

$$\lim_{x \to \infty} \frac{d}{d(\ln x)} f(x) = -(1 - \nu) + x \ln(\theta). \tag{4.16}$$

If $\theta = 1$, then as $x \to \infty$,

$$f(x) \sim \frac{1}{x^{1-\nu}}, \tag{4.17}$$

and ν has to be negative (since $f(x)$ has to yield a normalization). With negative ν and $\alpha = -\nu$, the GIGP distribution is from the class of Zipfian distributions. If $\theta = 1$ and $\nu < 2$, the GIGP distribution is Gaussian. If $\theta = 1$ and $\nu > 2$, the GIGP distribution is non-Gaussian. If $\theta < 1$, the GIGP distribution is a Gaussian non-Zipfian distribution. When $\beta = 0$ and $\nu = 0$, the GIGP distribution is reduced to the logarithmic distribution. Finally, when $\beta = 0$ and $\nu = 0$, the GIGP distribution is reduced to the negative binomial distribution.

The generalized Waring distribution and its particular cases will be much discussed in the next chapter. The generalized Waring distribution can be written in different mathematical forms. The form that expresses the distribution through the gamma and beta functions is

$$f(x) = \frac{\Gamma(a+c)}{B(a,b)\Gamma(c)} \frac{\Gamma(x+c)\Gamma(x+b)}{\Gamma(x+a+b+c)} \frac{1}{x!}. \tag{4.18}$$

The asymptotic behavior of this distribution as $x \to \infty$ is

$$f(x) \sim \frac{1}{x^{1+a}}. \tag{4.19}$$

Thus the generalized Waring distribution is a Zipfian distribution, and $\alpha = a$. if $a < 2$, the distribution is non-Gaussian. If $\alpha > 2$, the distribution is Gaussian.

In practice, one has to work with samples and calculate the moments of the corresponding distributions on the basis of the available samples. Thus the researcher has to observe the growth of the corresponding moments with increasing sample size N. In other words, one has to check the dependence of the mean and variance on N. If the dependence is negligible, then the corresponding population with large probability is a Gaussian one. If a dependence exists, then with large probability, the corresponding population is non-Gaussian.

4.7 Frequency Approach. Law of Lotka for Scientific Publications

The databases of scientific publications are an important final result of the activities of research organizations. And the production of research publications can be highly skewed. This means that in many research fields, a small number of highly productive researchers may be responsible for a significant percentage of all publications in the field.

Alfred Lotka (the same Lotka who is famous for the Lotka–Volterra model in populations dynamics [48–50]) investigated the database of the journal *Chemical*

Abstracts [22] and counted the number of scientists who wrote 1, 2, …, i_{max} papers. Lotka obtained the following relationship:

$$N_i = \frac{N_1}{i^2}, \tag{4.20}$$

where

- N_1: number of scientists who wrote one paper;
- N_i: number of scientists who wrote i papers.

Let us note that the law of Lotka doesn't consider the case $N_0 = 0$. This case may be considered on the basis of the Price distribution [51], which will be discussed in the next chapter within the scope of the discussion of the more general Waring distribution.

One may consider two variants of the law of Lotka [52–59] based on (4.20): a variant for the case of infinite productivity of the most productive scientist and a variant for the case of finite scientific productivity of the most productive scientist. Below we shall consider these two variants.

4.7.1 *Presence of Extremely Productive Scientists:* $i_{max} \rightarrow \infty$

Let N^* be the number of all scientists. Then we can introduce the proportions of the scientists who wrote i papers as $p_i = \frac{N_i}{N^*}$. From (4.20), we have

$$N^* = \sum_{i=1}^{i_{max}} N_i \approx \sum_{i=1}^{\infty} N_i = N_1 \sum_{i=1}^{\infty} \frac{1}{i^2} = N_1 \frac{\pi^2}{6}. \tag{4.21}$$

Then

$$p_i = \frac{N_i}{N^*} = \frac{N_1/i^2}{N_1/\frac{\pi^2}{6}} = \frac{6}{\pi^2} \frac{1}{i^2} \approx \frac{0.608}{i^2} \approx \frac{0.6}{i^2}, \tag{4.22}$$

where $\sum_{i=1}^{\infty} p_i = 1$. Equation (4.22) presents the law of Lotka:

> **The proportion of scientists who wrote i publications is inversely proportional to i^2 (the square of the number of publications).**

Let us stress that in order to investigate whether the law of Lotka is present in some database of scientific publications, we have to be sure that this database is large enough. Two additional remarks are in order here.

Remark 1 If we set $i = 1$ in the law of Lotka, we obtain that the minimally productive researchers (who wrote just one paper) constitute (at least) 60 % of the population of researchers. Then in a research organization, we can expect to find many researchers who have written a small number of papers (for a variety of reasons) and a small number of highly productive researchers.

Remark 2 In the general case, the exponent of the law of Lotka can be different from 2.

Another form of the law is

$$p_i = \frac{p_1}{i^{1+\alpha}}; \quad p_1 = \frac{1}{\zeta(1+\alpha)}, \qquad (4.23)$$

where α is the characteristic exponent of the law, and $\zeta(1+\alpha)$ is the Riemann zeta function: $\left(\zeta(\mu) = \sum_{i=1}^{\infty} \frac{1}{i^{\mu}}\right)$. If $\alpha = 1$, then the exponent in the law of Lotka is 2. The form of the law of Lotka (4.23) is similar to the *differential frequency form* of the Zipf distribution:

$$p(x) = \frac{C}{x^{1+\alpha}}, \ 0 \le \alpha < \infty, \qquad (4.24)$$

where C and α are parameters of the distribution. The Zipf distribution will be much discussed below. Let us note here that the *integral frequency form* of the Zipf distribution is

$$P(x) = \frac{C}{\alpha N} \left(\frac{1}{x_0^{\alpha}} - \frac{1}{x^{\alpha}}\right), \qquad (4.25)$$

where N, x_0, α, and C are parameters of the distribution.

The law of Lotka has been much discussed in connection with data sets for the publication activities of different categories of researchers [60, 61].

4.7.2 i_{max} Finite: The Most Productive Scientist Has Finite Productivity. Scientific Elite According to Price

The productivity of scientists is finite: $i_{max} \ne \infty$. In order to account for this, we have to set corrections to the above formulas. As we shall see, these corrections are small, and because of this, one often uses the formulas derived on the basis of the assumption of infinite productivity of the most productive researcher.

The finite productivity corrections will be based on the relationship [62]

$$\sum_{k=1}^{i_{max}} \frac{1}{k^2} \approx \frac{\pi^2}{6} - \frac{1}{i_{max}}. \qquad (4.26)$$

On the basis of this relation, the correction for the relationship (4.21) between the number of all researchers N^* and the number of researchers who have published one paper N_1 becomes

$$N^* = N_1 \left(\frac{\pi^2}{6} - \frac{1}{i_{max}} \right), \tag{4.27}$$

and the finite-size productivity correction of the proportion of researchers who have i publications becomes

$$p_i = \frac{N_i}{N^*} = \frac{6 i_{max}}{i^2 (\pi^2 i_{max} - 6)}. \tag{4.28}$$

Price defined the scientific elite as those researchers who have more than m publications, where m is such a number that the researchers who wrote more than m publications (the elite) possess the half the total number of publications of the group of researchers.

The result for the elite will be obtained on the basis of the following approximate relationship:

$$\sum_{i=1}^{i_{max}} \frac{1}{i} \approx \ln(n) + C_E, \tag{4.29}$$

where $C_E = 0.577 \ldots$ is Euler's constant.

The number of publications of the subgroup of researchers in which every researcher has i publications is $P(i) = iN_i$. The entire group of researchers obeys Lotka's law for scientific production. Then

$$P(i) = iN_i = i\frac{N_i}{i^2} = \frac{N_1}{i}. \tag{4.30}$$

Then half the total number of publications of the group of researchers is

$$\frac{1}{2}\sum_{i=1}^{i_{max}} P(i) = \frac{1}{2}\sum_{i=1}^{i_{max}} \frac{N_1}{i} \approx \frac{1}{2}N_1[\ln(i_{max}) + C_E]. \tag{4.31}$$

The number of researchers who have more than m publications is

$$\sum_{i=m}^{i_{max}} \frac{N_1}{i} = \sum_{i=1}^{i_{max}} \frac{N_1}{i} - \sum_{i=1}^{m} \frac{N_1}{i} \approx N_1[\ln(i_{max}) - \ln(i)]. \tag{4.32}$$

From (4.31) and (4.32), one obtains

$$m = \exp\left(-\frac{C_E}{2}\right) \sqrt{i_{max}} \approx 0.749\sqrt{i_{max}}. \tag{4.33}$$

> **Hence if the group of researchers have publications that obey the law of Lotka, then according to Price, the scientific elite consists of the researchers who have between $0.749\sqrt{i_{max}}$ and i_{max} publications**.

What is the size of this elite?
The number of elite scientists is

$$N_e = \sum_{i=m}^{i_{max}} \frac{N_1}{i^2} \approx N_1 \left(\frac{1}{m} - \frac{1}{i_{max}}\right). \tag{4.34}$$

The total number of scientists is given by (4.27). Thus the size of the elite of Price is

$$S_e = \frac{N_e}{N^*} = \frac{\pi(i_{max} - m)}{m(6i_{max} - \pi)}. \tag{4.35}$$

For the case of large maximum productivity i_{max},

$$S_e \approx \frac{\pi}{6m} = \frac{\pi}{6 \times 0.749\sqrt{i_{max}}} \approx \frac{0.812}{\sqrt{i_{max}}}. \tag{4.36}$$

Let $i_{max} = 250$. Then the size of the corresponding elite will be approximately 5 % of the size of the group of scientists. The research topic connected to scientific elites enjoys significant current interest, and that interest is very high especially for the study of highly cited researchers and publications [63–67].

4.7.3   The Exponent α as a Measure of Inequality. Concentration–Dispersion Effect. Ortega Hypothesis

According to the law of Lotka, the distribution of scientific production (the number of written publications) in a large enough group of researchers is determined by three parameters:

1. p_1: the percentage of minimally productive researchers;
2. i_{max}: the maximum productivity of a researcher from the group;
3. α: the exponent in the power law of Lotka.

If we fix one of the parameters, we can study the significance of one of the other parameters as a function of the third parameter. We are interested in the parameter α. Thus we fix i_{max} and discuss the relationship between α and p_1. From (4.23), one obtains

$$\frac{\partial p_1}{\partial \alpha} > 0, \tag{4.37}$$

which means that when α increases, the number of not very productive researchers increases too. At the same time, i_{max} = const, i.e., there is at least one highly productive researcher, but the number of highly productive researchers decreases with increasing α.

In other words, α is a measure of the stratification in a group of researchers with respect to the production characteristic called "number of published papers." And as α becomes larger, this stratification increases: there are more and more not very productive researchers and a smaller and smaller number of highly productive researchers.

The above stratification is one example of the concentration–dispersion effect.

Concentration–dispersion effect:
Two processes are simultaneously observed in organizations governed by hyperbolic laws: the concentration of units in a small number of components (formation of an elite) and dispersion of the rest of the units to many components of an organization.

The concentration–dispersion effect applied to our group of researchers means that there exists a small group of researchers that produce large number of publications, and there exists a large group of researchers who have only few publications each. In other words, we have to expect that most of the researchers will be not highly productive and that there will be small number of highly productive researchers. This doesn't mean that the research in the corresponding institution or country is not well organized. The periphery of low-productive researchers is a necessary part of the core–periphery structure, whereby the core contains a small number of highly productive researchers. One cannot try to eliminate the periphery without affecting the core. *The periphery contributes to the high productivity of the core.*

Social stratification [68–70] may arise in a research field as a consequence of the concentration–dispersion effect. A phenomenon similar to the concentration–dispersion effect may be observed also on the level of scientific fields (the few of them that are current attract many citations, and the other fields attract a much smaller number of citations).

A hypothesis called the *Ortega hypothesis* [71–80] is closely connected to the concentration–dispersion effect. Ortega suggests the following:

The work of the average scientists on unambiguous projects is very important for the advance of the science. The work of these scientists leads to minor contributions but without these minor discoveries by the mass of scientists the breakthroughs of the truly inspired scientists will be not possible [81].

4.7.4   The Continuous Limit: From the Law of Lotka to the Distribution of Pareto. Pareto II Distribution

If the number of researchers in the group is very large and the number of papers they have published is very large, too, then one can use a continuous approximation, whereby the number of publications $x(t)$ is a function of t (x is no longer necessarily a natural number).

> The continuous version of the law of Lotka is the distribution of (the law of) Pareto.

The distribution of Pareto [82, 83] is

$$p(x) = \frac{\alpha}{x_0} \left(\frac{x_0}{x}\right)^{1+\alpha}, \tag{4.38}$$

where

- $p(x)$: density of distribution of researchers;
- x_0: the minimum number of papers of researchers from the studied large group of researchers ($x_0 \leq x \leq \infty$).
- $\alpha > 0$

The law of Pareto can be obtained on the basis of two assumptions:

1. The time the researchers work on problems in some research area differs among the researchers from the group and is given by the distribution $p(t) = \nu \exp(-\nu t)$; (ν: parameter).
2. The number of publications of the researchers grows proportionally to the number of already written publications (more experience means a shorter time for writing a new publication): $dx/dt = \lambda x \rightarrow x(t) = x_0 \exp(\lambda t)$ (λ is a parameter; x_0 is the number of publications at the initial time t_0).

From the second assumption, $t = \frac{1}{\lambda} \ln \left(\frac{x}{x_0}\right)$. The substitution of this in the relationship for $p(t)$ from the first assumption leads to (4.38) with $\mu = \frac{\lambda}{1-\lambda x_0}$ and $\alpha = \frac{\mu}{\lambda} - 1$.

The Pareto distribution has a shortcoming that can be eliminated by the use of the Pareto II distribution. Let us discuss this in detail.

In the general Pareto distribution (4.38) above, x_0 is a scaling parameter. Let us define the standard Pareto distribution as

$$p_s = \frac{\alpha}{x^{\alpha+1}}. \tag{4.39}$$

Then if the random variable Y has standard Pareto distribution (4.39), the random variable $x_0 Y$ ($x_0 > 0$) has the general Pareto distribution (4.38).

The tail distribution function of the standard Pareto distribution and of the general Pareto distribution is (we assume $x > 1$)

$$P(Y > x) = \int_x^\infty dz \frac{\alpha}{z^{\alpha+1}} = \frac{1}{x^\alpha}. \tag{4.40}$$

Equation (4.40) shows very clearly a drawback of the standard Pareto distribution: *the smallest allowed value of x is* 1. In many distributions connected to science dynamics, however, values smaller than 1 are possible (one example is the value 0). In order to solve this problem, one may use the Pareto II distribution, which is obtained as follows [84, 85]: if Y is a random variable that has a standard Pareto distribution (4.39), then the random variable $X = \beta(Y - 1)$ has the Pareto II distribution

$$f_X(x) = \frac{\alpha\beta^\alpha}{(x + \beta)^{\alpha+1}}, \quad x \geq 0. \tag{4.41}$$

The tail distribution of the Pareto II distribution is

$$\Psi_X(x) = \left(\frac{\beta}{x + \beta}\right)^\alpha, \quad x \geq 0. \tag{4.42}$$

As one can see, the Pareto II distribution and its tail distribution are heavy-tailed: for large x, we have $f_X \propto 1/x^{\alpha+1}$; $\Psi_X \propto 1/x^\alpha$.

The Pareto II distribution can be adapted for variables in the interval $(1, \infty)$ by a simple shift $W = X + 1$. If the random variable X has Pareto II distribution, then the random variable W has the shifted Pareto II distribution

$$f_W(x) = \frac{\alpha\beta^\alpha}{(x + \beta - 1)^{1+\alpha}} \tag{4.43}$$

and tail function

$$P(W > x) = P(X > x - 1) = \left(\frac{\beta}{x + \beta - 1}\right)^\alpha. \tag{4.44}$$

Finally, the moments of the Pareto II distribution exist up to order $n < \alpha$, and the expected value of X^n is

$$E[X^n] = \beta^n n! \frac{\Gamma(\alpha - n)}{\Gamma(\alpha)}, \tag{4.45}$$

where $\Gamma(x)$ is the gamma function.

4.8   Rank Approach

4.8.1   Law of Zipf

The law of Zipf can be obtained from the law of Lotka as follows. The number of researchers who have at least i publications is

$$r_i = \sum_{k=i}^{i_{max}} N_k.$$ (4.46)

From the law of Lotka (4.23), we have $N_k = N_1 \frac{1}{k^{1+\alpha}}$. The substitution of the last relationship in (4.46) and letting $i_{max} \to \infty$ leads to

$$r_i = N_i \sum_{k=i}^{\infty} \frac{1}{k^{1+\alpha}} \approx \frac{N_1}{\alpha} \frac{1}{i^{\alpha}}.$$ (4.47)

Characteristics of r_i:
The number r_i is called the rank. According to (4.46), r_i is the characteristic of the number (in an ordered list) of researchers that have i publications.

Let us assume for simplicity that in the studied group we have researchers with different numbers of publications. Then the rank of the sole most productive researcher will be 1. If we take the number of publications of the second most productive researcher, then the number of researchers that have publications greater than or equal to the number of publications of the second most productive researcher will be 2 (and these are the most productive researcher and the second most productive researcher). Thus the rank of the second most productive researcher will be 2. The third most productive researcher will have rank 3, etc.

From (4.47), one obtains

$$i_r = \frac{B}{r^{\beta}},$$ (4.48)

where

$$B = \left(\frac{N_1}{\alpha}\right)^{1/\alpha} ; \quad \beta = \frac{1}{\alpha}.$$

Equation (4.48) with $\alpha = 1$ is called the law of Zipf [24, 86, 87].

4.8.2 Zipf–Mandelbrot Law

The Zipf–Mandelbrot law is obtained when we drop the assumption of infinite productivity of the most productive researcher, $i_{max} = \infty$, and instead of this assume finite productivity i_{max}. Then the result analogous to (4.47) is

$$r_k \approx \frac{N_1}{\alpha}\left(\frac{1}{i^\alpha} - \frac{1}{i_{max}^\alpha}\right). \tag{4.49}$$

From (4.49), one obtains the following rank distribution:

$$i_r = \frac{A}{(r+B)^\gamma}, \tag{4.50}$$

where

$$A = (N_1/\alpha)^{1/\alpha}; \quad B = [N_1/(i_{max}^\alpha \alpha)]; \quad \gamma = 1/\alpha,$$

which is called the Zipf–Mandelbrot law [88, 89].

If we set in (4.50) the value of α from the law of Lotka ($\alpha = 1$) and if we let $i_{max} \to \infty$, we shall obtain the law of Zipf (4.48).[2]

Equation (4.50) gives the differential rank form of the Zipf–Mandelbrot distribution. The integral rank form of the Zipf–Mandelbrot distribution is

$$R(r) = A\ln\left(\frac{r+B}{1+B}\right), \quad \gamma = 1 \tag{4.51}$$

and

$$R(r) = \frac{A}{\gamma - 1}\left[\frac{1}{(1+B)^{\gamma-1}} - \frac{1}{(r+B)^{\gamma-1}}\right], \quad \gamma \neq 1. \tag{4.52}$$

The Zipf–Mandelbrot law is much used not only in scientometrics but also in physics, applied mathematics, etc. [90–94]. Because of this, the practical aspects of fitting and testing this law are of great interest to researchers. A discussion of these aspects is provided in [95].

---

[2]Note that the Zipf law and the Zipf–Mandelbrot law give the rank of the researchers (or in general of some source of information) only approximately. The reason for this is that for example, there can be several researchers with the same production. If there are no researchers with the same production, then the laws are again approximate, because we silently changed the sums to integrals in order to obtain the approximate relationships for the final form of the laws.

4.8.3 Law of Bradford for Scientific Journals

The following simple classification of the scientific journals (with respect to the articles devoted to some scientific area) can be made:

1. **Core journals**: these are specialized in the corresponding area and contain many articles discussing different research questions from the area.
2. **Intermediate group of journals**: usually, these are journals devoted to closely related scientific areas and containing a certain number of articles on the scientific area being studied.
3. **Periphery journals**: Journals containing articles from other scientific areas and some articles from the studied scientific area.

One possible explanation for the appearance of such groups of journals is as follows. Every researcher tries to publish his/her manuscripts in the best journals. The number of pages of these journals is limited, however. Thus researchers have to publish in other journals as well. These other journals can be close to the research area, but some of them can be quite far from the area of research in the scientific field of interest.

Bradford applied the following approach to the ranked sources of information (journals) [96]. He separated them into groups containing sources of the same production: the journals were separated into a group of journals containing one paper on the studied research topic, then a group of journals containing two papers, etc. In doing so, Bradford obtained empirically the following law.

> **Law of Bradford**:
> *The journals ranked with respect to the number of articles on a scientific topic can be separated within groups of journals, each group containing the same total number of articles. Then the relationship between the numbers of journals in each group is*
> $$N_1 : N_2 : N_3 : \ldots = 1 : q : q^2 : \ldots,\qquad (4.53)$$
> *where $q > 1$ is some parameter (can be different for different research topics).*

The law of Bradford can be written as follows:

$$R(n) = k \ln\left(\frac{n}{n_0} + 1\right) \to k \ln n \quad \text{for large } \frac{n}{n_0}, \qquad (4.54)$$

where

- $R(n)$: the total number of papers in the first n journals from the highest ranked groups of journals.
- k: parameter depending on the number of papers in each group of journals and on the number q (for more information, see below).

- n_0: parameter depending on the number of journals in the group of highest ranking and on the number q (for more information, see below).

Equation (4.54) is obtained as follows. The number of journals in the L highest ranked groups of journals is $n = \sum_{i=1}^{L} n_i = n_1 \frac{q^L - 1}{q - 1}$ (this comes from the geometric progression in (4.53)). Let n^* be the number of journals in each of the L groups. Then the total number of journals is $R = Ln^*$, and $L = R/n^*$. Substituting this relationship for L into the above relationship for n, we obtain the law of Bradford, where

- $k = \frac{n^*}{\ln q}$,
- $n_0 = \frac{n_0}{q - 1}$.

In other words, the number of articles on a topic from a particular research area in the highest ranked n journals (if n is large) increases as the logarithm of the number of such journals.

With some algebra, one can obtain the following form of the law of Bradford:

$$R(n) \approx N_1 \ln n, \qquad (4.55)$$

where N_1 is the number of researchers who have written one or more articles on topics from the studied research area. This interesting relationship connects the number of researchers working on the research area of interest, the number of highest ranked journals in this area, and the number of published papers in these journals.

Let us note the following: *the law of Bradford is correct if the value of the exponent α in the law of Lotka is close to* 1. Such values are often present in practical situations, but there can be cases in which α can be significantly different from 1. Thus the Bradford law (as well as the other laws discussed above) has to be applied very carefully [97–99].

Let us stress the following.

> *The strongly positive feature of the laws discussed in this chapter is that they give us an orientation in the complex world of scientific structures, systems, and processes. For example, the frequent occurrence of $\alpha \approx 1$ is evidence of some kind of structure of the organization of science.*

Bradford's law may be used for obtaining information about the degree of inequality in scientific and technology between developed and developing nations and makes it possible to group them into three classes (core, middle, and periphery class) with respect to their science and technology self-reliance [100]. Bradford's law is observed also in the area of expending on research and development of firms and in processes of concentration of research and development [101].

Bradford's law describes the distribution of articles in a single discipline over the various journals. As some of the journals have become more and more interdisciplinary, this has complicated the conditions for the validity of the Bradford law

[102]. Bradford's law can depend on the stage of development of the corresponding scientific field. Thus the Bradford law can change over time [103].

The Bradford distribution can be connected to the Leimkuhler curve and to the index of Gini [104, 105]. In order to show the relationship between the Lorenz curve (much discussed in the previous chapter) and the Leimkuhler curve, we shall consider a population of N journals. For each journal, we consider a number of references (these references can be papers from a research area of interest). We assume that the numbers of references form a random variable X. Let us define:

- $F(j) = P(X \leq j)$;
- $\overline{F}(j) = \frac{r(j)}{N}$, where $r(j)$ is the rank of the journal carrying j references (i.e., the number of journals carrying at least j references).
- $\mu = \frac{M}{N}$, where M is the total number of references carried by the set of N journals being studied.

Now let

$$\Psi(j) = \sum_{i \geq j} \frac{iP(x = i)}{\mu}, \quad i = 1, 2, \ldots, \tag{4.56}$$

and

$$\Phi(j) = \sum_{i \leq j} \frac{iP(x = i)}{\mu}, \quad i = 1, 2, \ldots. \tag{4.57}$$

On the basis of the above definitions, we can define the Lorenz curve and the Leimkuhler curve as follows:

- **Lorenz curve**: the set of points $(F(j), \Phi(j))$;
- **Leimkuhler curve**: the set of points $(\overline{F}(j), \Psi(j))$.

The connection between the two curves becomes clear when one realizes that

$$\overline{F}(j) = 1 - F(j); \quad \Psi(j) = 1 - \Phi(j). \tag{4.58}$$

Hence if one can construct the Lorenz curve, then the construction of the Leimkuhler curve is an easy task.

4.9 Matthew Effect in Science

*For unto every one that hath shall be given, and
he shall have abundance: but from him that
hath not shall be taken even that which he hath.*
The Gospel of Matthew, Matthew 25:29

> **Matthew effect**: *a term for the phenomenon "the rich get richer and the poor get poorer" or "success breeds success."*

The term "Matthew effect" was introduced by Merton [106, 107] in order to give a name to a mechanism that increases the visibility of contributions to science by eminent researchers and reduces the visibility of comparable contributions by less well known authors. Merton assumed that a contribution would probably enjoy greater visibility when it was made by a scientist of higher eminence.

The Matthew effect helps the rapid diffusion of publications of eminent scientists [108] and especially of their publications that are not of top quality. Such papers written by high-ranking scientists are more likely to be widely diffused early than are papers of the same quality written by low-ranking authors.

The Matthew effect is observed at different scales of science dynamics. For example, there is a Matthew effect for countries: papers with authors who are from some countries get more citations than expected at the cost of others [109–113]. There exists a Matthew effect for journals (papers from more prominent journals are more frequently cited at the expense of papers from other journals) and even a Matthew effect for papers in one journal (the papers of authors from some nations are more cited than the papers by authors from other nations) [114–117]. The Matthew effect exists even with respect to the scientific centers that produce winners of scientific degrees and awards [118] as well as in the peer review process [119].

A measure of the characteristics of scientific systems connected to the Matthew effect is the Matthew index. It is defined as follows:

$$M = \frac{O - E}{E},\tag{4.59}$$

where

- O: observed number of items (say citations);
- E: expected number of items;

The Matthew index can be made more complicated in order to account for geographic areas [118] (e.g., in order to study the Matthew effect for academicians elected by the Chinese Academy of Sciences):

$$M_{ij} = \frac{O_{ij} - E_j}{E_j},\tag{4.60}$$

where

- O_{ij}: number of items from region i for year j;
- E_j: expected average number of items per region for year j.

Other characteristics of the Matthew effect can be studied by indexes of concentration discussed in Chap. 4 as well as by means of power law tests [120].

Another effect can be connected to Saint Matthew (second Matthew effect or invitation paradox [121]). This effect is connected to the fact that publishing in journals with a high impact factor does not imply a high number of citations, but offers a chance only: "For many are called, but few are chosen" (Matthew 22:14). Many papers published in journals of relatively high impact factor will be cited less frequently than the average, and relatively few papers obtain a high number of citations. This is not unexpected: if some papers are cited more than the expectation on the basis of the impact factor (which is an averaged quantity), then many papers will be cited less frequently than the expectation on the basis of the impact factor.

4.10 Additional Remarks on the Relationships Among Statistical Laws

In this chapter we have discussed the most famous statistical laws connected to bibliometrics and scientometrics. Bookstein [122] discusses the possibility that in spite of marked differences in their appearance, almost all statistical laws discussed in this chapter are variants of a single distribution. In several more words, Bookstein considers these statistical laws as differing manifestations of a single regularity, which he calls the *informetric law*. The basis for such a point of view is that the regularities (statistical laws) describe a population of discrete entities: researchers, journals, words, businessmen, etc., and each of these entities is producing something over a timelike variable (have some yield): researchers publish articles, articles occur in journals covering some scientific discipline, businessmen earn money, etc. Thus many of the statistical laws describe, in different ways, the same type of data: yields as distributed over a population of items.

Let us consider the above from the point of view of mathematics. Let us first write the classical statistical laws discussed above by means of a unified notation. Thus the law of Bradford may be written as

$$N_n = k^n N_0, \tag{4.61}$$

where k is a constant (equal to the constant q above in the text); N_0 and N_n are connected to the construction of Bradford: he formed a core of journals of central interest to the discipline, and then he formed rings of successively less productive journals, so that each ring contained the same number of relevant articles as the core. The number of journals in a ring divided by the number of journals in the preceding ring was approximately a constant k. Then N_0 is the number of journals in the core, and N_n is the number of journals from the nth ring. The Leimkuhler version of the Bradford law can be written as

$$Y = A \ln(1 + BN), \tag{4.62}$$

where the journals are ranked in decreasing order with respect to the productivity for the studied research discipline, and N is the number of journals required to yield Y articles. Here A and B are appropriate constants. Equation (4.62) (known also as a Leimkuhler distribution) can be written also as

$$N = A^*[\exp(B^*Y) - 1], \tag{4.63}$$

where A^* and B^* are constants.

Lotka's law was for the number f of researchers (chemists in the original version of the law) producing y articles,

$$f = \frac{A}{y^\alpha}, \tag{4.64}$$

where A is an appropriate constant and α is a constant approximately equal to 2. Zipf's law for the frequency y of word occurrence in natural text when the words are ranked according to the number of occurrences in the text is

$$ry = A, \tag{4.65}$$

where r is the rank of the word, y is the frequency (yield) of the word, and A is an appropriate constant. The Zipf–Mandelbrot law is written as

$$y = \frac{A}{(1 + Br)^\alpha}, \tag{4.66}$$

where A and B are appropriate constants.

Let us now briefly discuss the relations between the statistical laws. In the previous section, we have shown the equivalence between the Bradford law (4.61) and the logarithmic (Leimkuhler) form of this law (4.62) (for more mathematical detail about this equivalence, see [122]). Above we have shown that the laws of Zipf and Zipf–Mandelbrot can be obtained from the law of Lotka. Let us now show that there is a relationship between the law of Lotka (4.64) and the Leimkuhler form of the law of Bradford (4.62). Let us denote the expected maximum yield of an item by y_0 (we use the general terminology described at the beginning of this section). Let us rank the items with respect to their yield. Then the cumulative yield Y up to the items of rank r (the items of rank r are assumed to have yield y) is

$$Y = \sum_{n=y}^{y_0} nf_n, \tag{4.67}$$

where f_n is the number of items having a yield of n (e.g., the number of researchers who are authors of n articles). Assuming that the relationship for f_n is given by the law of Lotka (4.64), $f_n = (y_0/n)^\alpha$, we obtain

$$Y = y_0^\alpha \sum n^{1-\alpha}. \tag{4.68}$$

The integral approximation of (4.68), $Y \approx y_0^\alpha \int\limits_{y-1/2}^{y_0+1/2} dx \, x^{1-\alpha}$, is as follows:

1. Case $\alpha = 2$:

$$Y \propto y_0^2 \ln \left(\frac{y_0 + 1/2}{y - 1/2} \right). \tag{4.69}$$

2. Case $\alpha \neq 2$:

$$Y \propto \frac{y_0^\alpha}{2 - \alpha} \left[\frac{1}{(y_0 - 1/2)^{\alpha-2}} - \frac{1}{(y - 1/2)^{\alpha-2}} \right]. \tag{4.70}$$

The rank r of the items of yield y in the presence of the law of Lotka is $\sum\limits_{x=y}^{y_0} (y_0/x)^\alpha$, which can be approximated as $\int_{y-1/2}^{y_0+1/2} dx \, (y_0/x)^\alpha$. Then (note that $\alpha \neq 1$)

$$r = \frac{y_0}{\alpha - 1} \left[\left(\frac{y_0}{y - 1/2} \right)^{\alpha-1} - \left(\frac{y_0}{y_0 + 1/2} \right)^{\alpha-1} \right]. \tag{4.71}$$

From (4.71), one obtains

$$\frac{1}{y - 1/2} = \left[\frac{(\alpha - 1)r}{y_0^\alpha} + \left(\frac{1}{y_0 + 1/2} \right)^{\alpha-1} \right]^{1/(\alpha-1)}, \tag{4.72}$$

and the substitution of this in (4.71) leads to

1. $\alpha = 2$:

$$Y \approx A[(B + Cr)^\alpha + 1], \tag{4.73}$$

2. $\alpha \neq 2$:

$$Y \approx A \ln(1 + Br), \tag{4.74}$$

where A and B are appropriate constants that can be easily calculated by the interested reader. In a similar way, one can obtain the Zipf law from the Leimkuhler version of the law of Bradford as well as the law of Pareto from the law of Lotka [122].

4.11   On Power Laws as Informetric Distributions

As we have noted at the beginning of the chapter, the laws connected to the processes studied by scientometrics, bibliometrics, and informetrics should be understood not literally, but as statements about probability distributions, or as statements about the corresponding expected values. Let us consider a population of objects and let each object of this population have integer yield y that can be measured. We can associate

another yield to each of the objects: the expected yield x. This expected yield may be not be an integer, and it may not be measurable. Let the number of objects (e.g., researchers) $f(x)$ having an expected yield x (e.g., publications) be proportional to a function $h(x)$ of x [123]:

$$f(x) = Ah(x), \tag{4.75}$$

where A is a constant (which may be set if we assume $h(1) = 1$). The relation between the expected yield x and the actually measured value is as follows. Let $p(n \mid x)$ be the probability that an object (researcher) with expected yield of x units (articles) actually has n units. Then the number of objects with n units will be proportional to

$$g(n) = \int dx \, p(n \mid x)h(x). \tag{4.76}$$

If $p(n \mid x)$ is sharply peaked at n near x, then $g(n) \approx h(n)$. Under the condition of sharp-peaked conditional probability, we have $f(n) \approx Ah(n)$ if the density of expectations is proportional to $h(x)$ (even x is a noninteger). *Thus instead of discrete values n for the units, one may work with continuous values x of the yield variable consistent with the expected value interpretation.* Bookstein [124] gives an example of the usefulness of this approach: if $h(x) = 1/x^2$ and $p(n \mid x)$ is a Poisson distribution, then the expected number of objects yielding n events (units) is $A/[n(n-1)]$, which for large values of n is approximately A/n^2.

After validation of the possibility of working with a continuous variable x instead of with the discrete variable n (and to obtain correct results), let us discuss the question of the form of the distribution $h(x)$ from (4.75) if we change the time interval from an interval in which every object produces the expected value of x units to an interval in which every object produces an expected value of sx units. We impose the following conditions:

1. Our law $h(x)$ has to be stable over such kinds of changes, i.e., the form of our distribution for the case of sx units will again be of the form h: $h(sx)$.
2. *The members of the population of objects produce units at a constant rate*.
3. The population of objects is stable (there are no entries and no exits of objects).

The above conditions lead to statistical laws in the form of power-law distributions. Let us show this.

The population of considered objects (scientists) produced x units (articles) in the first period and $x' = sx$ units in the second period. The number of objects having expected value $x' + \Delta$ in the second period will have expected values between $x'/s + \Delta/s = x + \Delta/s$ in the first period. We know the number of these objects for the first period: on the basis of (4.75), this number is $f(x)(\Delta/s) = Ah(x)(\Delta/x) = Q$. And then the number of objects that have produced Δ units in the second period is $Ah(x)/s$ (i.e., Q/Δ).

The form of the distribution should be stable, i.e., it should remain a constant multiplied by the function h of the expected number of produced units. Then

$$A'h(sx) = A\frac{h(x)}{s}. \tag{4.77}$$

From the condition $h(1) = 1$ and setting $x = 1$, we obtain

$$A'h(s) = \frac{A}{s}. \tag{4.78}$$

From (4.77) and (4.78), we obtain

$$h(sx) = h(s)h(x). \tag{4.79}$$

Equation (4.79) determines the form of the function $h(x)$. Taking into account that $x + \Delta = x(1 + \Delta/x)$, we obtain from (4.79)

$$h(x + \Delta) = h(x)h(1 + \Delta/x), \tag{4.80}$$

and this leads us to the relationship

$$\frac{h(x + \Delta) - h(x)}{\Delta} = \frac{h(x)}{x}\frac{h(1 + \Delta/x) - h(1)}{\Delta/x}. \tag{4.81}$$

Assuming that $(d/dx)[h(1)]$ exists, we obtain by letting $\Delta \to 0$,

$$\frac{dh(x)}{dx} = \frac{h(x)}{x}\frac{dh(1)}{dx}, \tag{4.82}$$

where dh/dx evaluated at $h = 1$ is a constant that we denote by A. Then

$$\frac{dh(x)}{dx} = A\frac{h(x)}{x}, \tag{4.83}$$

which has a power-law function as general solution, i.e.,

$$h(x) = Ax^{-\alpha}, \tag{4.84}$$

where α may be an arbitrary constant (but in practice, $\alpha > 0$, since $h(x)$ is connected to a statistical distribution).

It is remarkable that the relationship (4.84) is present even if one relaxes the requirement for stability of the population of objects, i.e., when objects (researchers) may enter and leave the population [125]. In more detail, the form of $h(x)$, namely $h(x) = Ax^{-\alpha}$, is maintained if the objects enter and leave the population at arbitrary rates, provided that *the distribution in yield production of the items entering and*

leaving the population is the same as that of those initially in the population. In addition, the above form for $h(x)$ is the only form for which this is true. The same form occurs if the rates of production are varying, i.e., if the objects don't generate items at a fixed rate [123]. In more detail, the rate of change may be arbitrary, and the condition is that a change in the rate affects all objects in the same way.

Thus the occurrence of a power-law relationship may be considered a sign of the inertia of productivity patterns. The stability of the power laws in bibliometrics (e.g., of the law of Lotka) is consistent with the recognition that the studied research discipline will experience slow periods and periods of acceleration. If these variations over time tend to influence all the members of the discipline in the same way, then the corresponding power law will be preserved.

Let us now discuss the relation between the power laws and the multiple authorship (several objects are coauthors of a unit). Lotka derived his law by giving full credit to the senior author and to him alone (i.e., nothing for the other coauthors). It can be shown [123] that if the power law $h(x) \propto 1/x^{\alpha}$ is valid for one accounting system for authorship, it will be valid for any other if certain regularities exist. In addition, this law will be unique in being invariant under changes in counting method for x the expected yield in published articles. *Then if one finds a $1/x^{\alpha}$ relation to describe productivity for the case in which a full publication credit is assigned to every author whose name appears on a paper, this will also be the case if we had assigned fractional authorship instead.*

Let us now add some mathematics to the statements above. The number of objects (researchers) that are expected to produce between x and $x + dx$ units (articles) in some time interval are (as above) $Ah(x)dx$, where A is a constant defined by the constraint $h(1) = 1$ and $h(x)$ describes the studied population of objects for *some basis of accounting*. The question is whether the form of $h(x)$ changes if we change the accounting system, e.g., to the accounting system we are currently using. We consider an object (researcher) that has produced N units (articles): n_1 as lone author, n_2 with 1 coauthor, n_3 with two coauthors, etc. Let us in general use an accounting system that assigns credit v_i for the ith unit(paper) of the object (author). Then the total number of units that will be assigned to the object of interest is

$$x = \sum_{i=1}^{N} = rN, \tag{4.85}$$

where r is defined by (4.85). On the basis of this system of accounting, we have $Ah(x)dx$ objects who are expected to yield between x and $x + dx$ units.

Now let us consider a different accounting system. From the point of view of this system, the object yields x' units. This can be written as

$$x' = \frac{r'}{r}x = \theta x; \quad \theta = r'/r, \tag{4.86}$$

where θ depends on the object and on the accounting system. Now we shall obtain an equation for θ starting from the question, How many objects (authors) will yield x' units (articles) in the new accounting system? The number of objects that yield x' units with respect to the new accounting system is equal to the number of objects that yield x'/θ units from the point of view of the old system of accounting. Taking into account that the number of objects that have values of θ between θ_0 and $\theta_0 + d\theta$ is $F(\theta)d\theta$ (where $F(\theta)$ is the probability density function of θ), we obtain that the function $A'h'(x')$ connected to the objects yielding x' units is

$$A'h'(x') = A \int d\theta \, F(\theta) \left[\frac{1}{\theta} h \left(\frac{x'}{\theta} \right) \right], \tag{4.87}$$

where the factor $1/\theta$ compensates for the change of size of dx before and after the transformation.

Now suppose that a change in the accounting system does not change the form of h, i.e.,

$$h'(x) = h(x). \tag{4.88}$$

The substitution of (4.88) in (4.87) leads to

$$A'h(x) = A \int d\theta \, F(\theta) \left[\frac{1}{\theta} h \left(\frac{x}{\theta} \right) \right]. \tag{4.89}$$

Taking into an account that $h(1) = 1$, we obtain

$$A' = A \int d\theta \, F(\theta) \left[\frac{1}{\theta} h \left(\frac{1}{\theta} \right) \right]. \tag{4.90}$$

The substitution of (4.90) in (4.89) leads to

$$h(x) \int d\theta \, F(\theta) \left[\frac{1}{\theta} h \left(\frac{1}{\theta} \right) \right] = \int d\theta \, F(\theta) \left[\frac{1}{\theta} h \left(\frac{x}{\theta} \right) \right]. \tag{4.91}$$

Given $F(\theta)$, (4.91) is an equation for $h(x)$. It is straightforward to check that $h(x) = 1/x^\alpha$ is a solution of this equation. Moreover, the power law is the only solution that satisfies all constraints imposed to the problem. *Then, we can conclude that if the above power law holds for one accounting method, it will hold for every other one in which the change in the typical amount of credit given to authors per paper may*

vary from author to author but does not depend strongly on how much the author published. Then if the objects are authors and units are articles [123]:

> the investigator is free to adopt any reasonable system of assigning credit, and can be confident that if power law isn't observed, it is not because he chose the wrong means of attributing articles to authors.

Finally let us note that if we have several classes of objects that yield units and the distribution of those units follows some power law, then if we for some reason do not distinguish between the classes of objects (e.g., between chemists and biologists), then the distribution of units connected to the class of all objects will be approximately a power law [123].

There are many models that lead to relationships connected to different aspects of research production. A large number of such models will be discussed in the next chapter.

References

1. D. de Solla Price, A general theory of bibliometric and other cumulative advantage processes. J. Am. Soc. Inf. Sci. **27**, 292–306 (1976)
2. L. Egghe, The exact place of Zipf's and Pareto's law amongst the classical informetric laws. Scientometrics **20**, 93–106 (1991)
3. M.G. Kendall, Natural law in the social sciences. J. R. Stat. Soc. **124**, 4–16 (1961)
4. B.C. Brookes, The derivation and application of the Bradford-Zipf Distribution. J. Documentation **24**, 247–265 (1968)
5. B.C. Brookes, Bradford's law and the bibliography of science. Nature **224**, 953–956 (1969)
6. R.A. Fairthorne, Empirical hyperbolic distribution (Bradford-Zipf-Mandelbrot) for bibliometric description and prediction. J. Documentation **25**, 319–349 (1969)
7. G. Scarrot, Will Zipf join Gauss? New Sci. **62**, 402–404 (1974)
8. J. Vlachy, Frequency distribution of scientific performance. A bibliography of Lotka's law and related phenomena. Scientometrics **1**, 109–131 (1978)
9. S.D. Haitun, Stationary scientometric distributions. Part I. Different approximations. Scientometrics **4**, 5–25 (1982)
10. L. Egghe, Zipfian and Lotkaian continuous concentration theory. J. Am. Soc. Inf. Sci. Technol. **56**, 935–945 (2005)
11. L. Egghe, Consequences of Lotka's law for the law of Bradford. J. Documentation **41**, 173–189 (1985)
12. L. Egghe, The power of power laws and an interpretation of Lotkaian infometric systems as self-similar fractals. J. Am. Soc. Inf. Sci. Technol. **56**, 669–675 (2005)
13. L. Egghe, Consequences of the Lotka's law in the case of fractional counting of authorship and first author counts. Math. Comput. Model. **18**, 63–77 (1993)
14. S.D. Haitun, Stationary scientometric distributions. Part II. Non-gaussian nature of scientific activities. Scientometrics **4**, 89–104 (1982)
15. S.D. Haitun, Stationary scientometric distributions. Part III. The role of the Zipf distribution. Scientometrics **4**, 181–194 (1982)
16. L. Egghe, R. Rousseau, Theory and practice of the shifted Lotka function. Scientometrics **91**, 295–301 (2012)

17. A.I. Yablonsky, On fundamental regularities of the distribution of scientific productivity. Scientometrics **2**, 3–34 (1980)
18. K. Prpic, The publication activity of young scientists: an empirical study. Scientometrics **49**, 453–490 (2000)
19. K. Prpic, Gender and productivity differentials in science. Scientometrics **55**, 27–58 (2002)
20. R.K. Toutkoushian, S.R. Porter, C. Danielson, P.R. Hollis, Using publications counts to measure an institutions;s research productivity. Res. High. Educ. **44**, 121–148 (2003)
21. S. Kyvik, Productivity differences, fields of learning, and Lotka's law. Scientometric **15**, 205–214 (1989)
22. A.J. Lotka, The frequency distribution of scientific productivity. J. Washington Acad. Sci. **16**, 317–323 (1926)
23. S.C. Bradford, Sources of information on specific subjects. Engineering **137**, 85–86 (1934)
24. G.K. Zipf, *Human Behavior and the Principle of Least Effort* (Addison-Weasley, Cambridge, MA, 1949)
25. D. de Solla Price, Little Science, Big Science (Columbia University Press, New York, 1963)
26. J.M. Russel, R. Rousseau, Bibliometrics and institutional evaluation, in Encyclopedia of Life Support Systems (EOLSS), ed. by R. Arvantis. Part 19.3: Science and Technology Policy (EOLSS Publishers, Oxford, UK, 2002), pp. 1–20
27. E. Garfield, *Citation indexing—its theory and application in science, technology, and humanities* (Wiley, New York, 1979)
28. E. Garfield, Journal inpact factor: a brief review. Can. Med. Assoc. J. **161**, 979–980 (1999)
29. B.K. Sen, Normalised impact factor. J. Documentation **48**, 318–325 (1992)
30. E. Garfield, The impact factor and using it correctly. Unfallchirurg **101**, 413–414 (1998)
31. S.J. Bensman, Garfield and the impact factor. Ann. Rev. Inf. Sci. Technol. **41**, 93–155 (2007)
32. A. Smith, M. Thelwall, Web impact factors for Australasian universities. Scientometrics **54**, 363–380 (2002)
33. M. Thelwall, Journal impact evaluation: a webometric perspective. Scientometrics **92**, 429–441 (2012)
34. P. Ingwersen, The calculation of web impact factors. J. Documentation **54**, 236–243 (1998)
35. L. Leydesdorff, T. Opthof, Scopus's source normalized impact per paper (SNIP) versus a journal impact factor based on fractional counting of citations. J. Am. Soc. Inf. Sci. Technol. **61**, 2365 – 2369 (2010)
36. S.D. Haitun, *Scientometrics: State and Perspectives* (Nauka, Moscow, 1983). (in Russian)
37. B.V. Gnedenko, On the theory of limit theorems for sums of independent random variables. Bull. Acad. Sci. URSS. Ser. Math. [Izvestia Akad. Nauk SSSR] **1939**, 643–647 (1939). (in Russian)
38. W. Doeblin, Sur l'ensemble de puissonces d'une loi de probabilitè. Studia Math. **9**, 71–96 (1940)
39. S.D. Haitun, *Quantitative Analysis of Social Phenomena* (URSS, Moscow, 2005). (in Russian)
40. N.H. Bingham, Regular variation and probability: the early years. J. Comput. Appl. Math. **200**, 357–363 (2007)
41. B.B. Mandelbrot, *Fractals and Scaling in Finance* (Springer, New York, 1997)
42. M. Falk, J. Hüsler, R. Reiss, *Laws of Small Numbers: Extremes and Rare Events* (Birkhäuser, Basel, 2011)
43. J.W. Lamperti, *Probability. A Survey of the Mathematical Theory* (Wiley, New York, 1996)
44. A.I. Yablonsky, Stable non-Gaussian distributions in scientometrics. Scientometrics **7**, 459–470 (1985)
45. N.L. Johnson, S. Kotz, N. Balakrishnan, *Continuous Univariate Distributions*, vol. I (Wiley, New York, 1994)
46. A. Papoulis, *Probability, Random Variables, and Stochastic Processes* (McGraw-Hill, New York, 1984)
47. S.D. Haitun, Criteria of Gaussian/non-Gaussian nature of distributions and populations, in Informetrics 89/90, ed. by L. Egghe, R. Rousseau (Elsevier, Amsterdam, 1990), pp. 149–161
48. A. Lotka, *Elements of Physical Biology* (Williams and Wilkins, Baltimore, 1925)

49. A.J. Lotka, *Analytical Theory of Biological Populations* (Plenum Press, New York, 1998)
50. H. Voos, Lotka nad information science. J. Am. Soc. Inf. Sci. **25**, 270–272 (1974)
51. W. Glänzel, A. Schubert, Price distribution. An exact formukation of Price's "square root law". Scientometrics **7**, 211–219 (1985)
52. L. Egghe, Relations between the continuous and the discrete Lotka power function. J. Am. Soc. Inf. Sci. Technol. **56**, 664–668 (2005)
53. R.C. Coile, Lotka's frequency distribution of scientific productivity. J. Am. Soc. Inf. Sci. **28**, 366–370 (1977)
54. L.J. Murphy, Lotka's law in the humanities? J. Am. Soc. Inf. Sci. **24**, 461–462 (1973)
55. T. Radhakrishnan, R. Kennzan, Lotka's law and computer science literature. J. Am. Soc. Inf. Sci. **30**, 51–54 (1979)
56. K.H. Chung, R.A.K. Cox, Patterns of productivity in the finance literature: a study of the bibliometric distributions. J. Finance **45**, 301–309 (1990)
57. N. Kumar, Applicability to Lotka's law to research productivity of council of scientific and industrial research (CSIR), India. Ann. Libr. Inf. Sci. **57**, 7–11 (2010)
58. M.I.M. Sobrino, A.I.P. Caldes, A.P. Guerrero, Lotka law applied to the scientific production of information science area. Braz. J. Inf. Sci. **2**, 16–30 (2008)
59. B.K. Sen, Lotka's Law: a viewpoint. Ann. Libr. Inf. Stud. **57**, 166–167 (2010)
60. M.L. Pao, Lotka's law: a testing procedure. Inf. Process. Manage. **21**, 305–320 (1985)
61. M.T. Kinnucan, D. Wolfram, Direct comparison of bibliometric models. Inf. Process. Manage. **26**, 777–790 (1990)
62. P.D. Allison, D. de Sola, Price, B.G. Griffith, J.M. Moravcsik, J.A. Stewart, Lotka's law: a problem in its interpretation and application. Soc. Stud. Sci. **6**, 269–276 (1976)
63. V. Lariviere, B. Macaluso, E. Archambault, Y. Gingras, Which scientific elites? On the concentration of research funds, publications and citations. Res. Eval. **19**, 45–53 (2010)
64. A.H. Goodall, Highly cited leaders and the performance of research universities. Res. Policy **38**, 1079–1092 (2009)
65. D.W. Asknes, G. Sivertsen, The effect of highly cited papers on national citation indicators. Scientometrics **59**, 213–224 (2004)
66. I. Podllubny, K. Kassayova, Towards a better list of citation superstars: compiling a multidisciplinary list of highly cited researchers. Res. Eval. **15**, 154–162 (2006)
67. C. Cao, *China's Scientific Elite* (RouthledgeCurzon, London, 2004)
68. J.R. Cole, S. Cole, *Social Stratification in Science* (The University of Chicago Press, Chicago, 1973)
69. M. Albrow, E. King (eds.), *Globalization, Knowledge and Society* (Sage, London, 1990)
70. X. Zhou, A nonparametric index of stratification. Sociol. Methodol. **42**, 365–389 (2012)
71. J. Ortega y Gasset, *The Revolt of the Masses* (Norton, New York, 1932)
72. M.H. Macroberts, B.R. Macroberts, Testing the Ortega hypothesis: facts and artifacts. Scientometrics **12**, 293–295 (1987)
73. A. Meadows, Ortega hypothesis. Scientometrics **12**, 315–316 (1987)
74. A. Nederhof, A. van Raan, Citation theory and the Ortega hypothesis. Scientometrics **12**, 325–328 (1987)
75. M. Oromaner, The Ortega hypothesis and influential articles in American sociology. Scientometrics **7**, 3–10 (1985)
76. S. Cole, J. Cole, Testing the Ortega hypothesis: milestone or millstone? Scientometrics **12**, 345–353 (1987)
77. H. Kretschmer, Measurement of social stratification. A contribution to the dispute on the Ortega hypothesis. Scientometrics **26**, 97–113 (1993)
78. H. Kretschmer, R. Müller, A contribution to the dispute on the Ortega hypothesis: connection between publication rate and stratification of scientists, tested by various methods. Scientometrics **18**, 43–56 (1990)
79. W. Snizek, A re-examination of the Ortega hypothesis: the Dutch case. Scientometrics **9**, 3–11 (1986)

80. L. Bornmann, F. de Moya Anegon, L. Leydesdorff, Do scientific advancements lean on the shoulders of giants? A bibliometric investigation of the Ortega hypothesis. PLOS One **5**, e.11344 (2010)
81. J.R. Cole, S. Cole, The Ortega hypothesis. Science **178**(4059), 368–375 (1972)
82. B.C. Arnold, Pareto and generalized Pareto distributions, in *Modeling Income Distributions and Lorenz Curves*, ed. by D. Chotikapanich (Springer, New York, 2008), pp. 119–145
83. M.E.J. Newman, Power laws, Pareto dostributions and Zipf's law. Contemp. Phys. **46**, 323–351 (2005)
84. Q.L. Burrell, Extending Lotkaian infometrics. Inf. Process. Manage. **44**, 1794–1807 (2008)
85. B.C. Arnold, *Pareto Distributions* (International Co-operative Publishing House, The University of Michigan, MI, 1983)
86. X. Gabaix, Zipf's law for cities: an explanation. Q. J. Econ. **114**, 739–767 (1999)
87. W. Li, Information theory. IEEE Trans. Inf. Theory **38**, 1842–1845 (1992)
88. B. Mandelbrot, *The Fractal Geometry of Nature* (Freeman, New York, 1982)
89. Z.K. Silagadze, Citations and the Zipf-Mandelbrot law. Complex Syst. **11**, 487–499 (1997)
90. M.A. Montemurro, Beyond the Zipf-Mandelbrot law in quantitative linguistic. Phys. A **300**, 567–578 (2001)
91. V.P. Maslov, The Zipf-Mandelbrot law: quantization and an application to the stock market. Russ. J. Math. Phys. **12**, 483–488 (2005)
92. S. Abe, N. Suzuki, Scale-free statistics of time interval between successive earthquakes. Phys. A **350**, 588–596 (2005)
93. P. Louridas, D. Spinellis, V. Vlachos, Power laws in software. ACM Trans. Softw. Eng. Methodol. **18**(1)(2) (2008)
94. B. Manaris, D. Vaughan, C. Wagner, J. Romero, R.B. Davis, Evolutionary music and the Zipf-Mandelbrot law: Developing fitness functions for pleasant music. Lect. Notes Comput. Sci. **2611**, 522–534 (2003)
95. J. Izsak, Some practical aspects of fitting and testing the Zipf-Mandelbrot model. A short essay. Scientometrics **67**, 107–120 (2006)
96. S.C. Bradford, Sources of information on specific subjects. J. Inf. Sci. **10**, 176–180 (1985)
97. S. von Ungern-Sternberg, Bradford's law in the context of information provision. Scientometrics **49**, 161–186 (2000)
98. K.C. Garg, P. Sharma, L. Sharma, Bradford's law in relation to the evolution of a field. A case study of solar power research. Scientometrics **27**, 145–156 (1993)
99. W. Chongde, W. Zhe, Evaluation of the models for Bradford's law. Scientometrics **42**, 89–95 (1998)
100. H. Eto, P.M. Candelaria, Applicability of the Bradford distribution to international science and technology indicators. Scientometrics **11**, 27–42 (1987)
101. H. Eto, Bradford law in R&D expending of firms and R&D concentration. Scientometrics **6**, 183–188 (1984)
102. A. Bookstein, Towards multi-disciplinary Bradford law. Scientometrics **30**, 353–361 (1994)
103. R. Wagner-Döbler, Time dependencies of Bradford distributions: structures of journal output in 20th-century logic and 19-th century mathematics. Scientometrics **39**, 231–252 (1997)
104. Q.L. Burrell, The Bradford distribution and the Gini index. Scientometrics **21**, 181–194 (1991)
105. F.F. Leimkuhler, The Bradford distribution. J. Documentation **23**, 197–207 (1967)
106. R.C. Merton, The Matthew effect in science. Science **159**(3810), 56–63 (1968)
107. R.K. Merton, The Matthew effect in science II: cumulative advantage and symbolism of intelectual property. ISIS: J. Hist. Sci. Soc. **79**, 606–623 (1988)
108. J.R. Cole, S. Cole, *Social Stratification in Science* (The University of Chicago Press, Chicago, 1973)
109. M. Bonitz, E. Bruckner, A. Scharnhorst, Characteristics and impact of the Matthew effect for countries. Scientometrics **40**, 407–422 (1997)
110. M. Bonitz, Ten years Matthew effect for countries. Scientometrics **64**, 375–379 (2005)
111. M. Bonitz, E. Bruckner, A. Scharnhorst, The micro-structure of the Matthew effect for countries, in *Proceedings of the Seventh Conference of the International Society for Scientometrics and Infometrics* (Universidad de Colima, Colima, Mexico, 1999), pp. 50–64

112. V. Lariviere, Y. Gingras, Brief communication: the impact factor's Matthew effect: a natural experiment in bibliometrics. J. Am. Soc. Inf. Sci. Technol. **61**, 424–427 (2010)
113. J. Wang, Unpacking the Matthew effect in citations. J. Infometrics **8**, 329–339 (2014)
114. M.H. Biglu, The influence of references per paper in the SCI to Impact Factors and the Matthew effect. Scientometrics **74**, 453–470 (2008)
115. M. Bonitz, E. Bruckner, A. Scharnhorst, The Matthew index—concentration patterns and Matthew core journals. Scientometrics **44**, 361–378 (1999)
116. M. Bonitz, A. Scharnhorst, Competition in science and the Matthew core journals. Scientometrics **51**, 37–54 (2001)
117. V. Pislyakov, E. Dyachenko, Citation expectations: are they realized? Study of the Matthew effect for Russian papers published abroad. Scientometrics **83**, 739–749 (2010)
118. X. Yang, X. Gu, Y. Wang, G. Hu, L. Tang, The Matthew effect in China's science: evidence from academicians of Chinese Academy of Sciences. Scientometrics **102**, 2089–2105 (2015)
119. F. Squazzoni, C. Candelli, Saint Matthew strikes again: an agent based model of the scientific community structure. J. Infometrics **6**, 265–275 (2012)
120. B. Birkmaier, K. Wohlrabe, The Matthew effects in economics reconsidered. J. Infometrics **8**, 880–889 (2014)
121. P. Vinkler, Introducing the contemporary contribution index for characterizing the recent relevant impact of journals, in *Proceedings of the International Conference of the International Society for Scientometrics and Informetrics*, ed. by D. Torres-Salinas, H.F. Moed (CINDOC-CSIC, Madrid, 2007), pp. 753–760
122. A. Bookstein, Informetric distributions, Part I: unified overview. J. Am. Soc. Inf. Sci. **41**, 368–375 (1990)
123. A. Bookstein, Informetric distributions, Part II: resilience to ambiguity. J. Am. Soc. Inf. Sci. **41**, 376–386 (1990)
124. A. Bookstein, The bibliometric distributions. Libr. Q. **46**, 416–423 (1976)
125. A. Bookstein, Patterns of scientific productivity and social change: a discussion of Lotka's law and bibliometric symmetry. J. Am. Soc. Inf. Sci. **28**, 206–210 (1977)