Research

Comparative Analysis of Author Productivity Concentration of Different Domains through Observing the NACSIS Database of Academic Conference Papers

著者生産性分布における集中度の分野間比較：NACSIS 学会発表データベースの分析を通して

Faculty of University Evaluation and Research,
National Institution for Academic Degrees and University Evaluation
大学評価・学位授与機構　評価研究部

YOSHIKANE, Fuyuki
芳　鐘　冬　樹

Abstract

Many studies have examined concentration of author productivity, using various concentration measures. However, most studies are irrelevant in that they do not consider the sample size dependency of measures. In this article, taking into consideration the sample size dependency, we compare five different domains in terms of two aspects, i.e., concentration in the whole productivity distribution and behavior of core authors in the domain.

1. Introduction

In this article, we compare five different domains in terms of concentration of author productivity, measured by the number of papers written by each author. Taking into consideration the sample size dependency of measures, we describe the characteristics of author productivity in terms of two aspects, i.e., concentration in the whole productivity distribution, and behavior of core authors in the domain.

The survey of actual states concerning such bibliometric phenomena as the publication of journals and the performance of researchers, which is based on bibliographical information sources, has been one of the main themes in library and information sciences. Various bibliometric measures have been proposed for various purposes and from various viewpoints (cf., Kishida, 1996). Among concentration measures for bibliometric distributions, Pratt's measure *C* (Pratt, 1977) is well-known. There are many studies that have applied *C* or other concentration measures (e.g., Egghe, 1988a; 1988b). However, most of these are irrelevant in that they do not consider the sample size dependency of measures. In our previous study (Yoshikane, Kageura, & Tsuji, 2003), we introduced two types of concentration, namely absolute concentration and relative concentration, and proposed a useful method for comparing author productivity concentration on the basis of different samples of different sizes.

This study applies the method observing the developmental profiles of measures for increasing sample size, which is shown in Yoshikane et al. (2003). Moreover, we focus on "relative" concentration (inequality) in author productivity distributions, and use two different measures of relative concentration. Using one measure, we examine the tendency of the whole domain, while using the other, we examine that of core researchers, whose behavior is said to be independent of that of the other researchers in the domain (cf., Pao, 1986; Yoshikane & Kageura, 1999). We aim to describe the characteristics of author productivity in greater detail in terms of these two aspects. The target domains in this study are information processing, electrical engineering, architecture, polymer science, and biochemistry. We chose these domains as representatives of computer sciences, electrical engineering, civil engineering, chemistry, and biology, respectively.

This article is organized as follows. In Section 2, we elucidate the concept of "concentration" to be evaluated in this study, and describe the characteristics of concentration measures. In Section 3, we discuss the peculiarity of author productivity data, that is, the sample size dependency of statistical measures. We then explain our samples for examining concentration of author productivity. Lastly, in Section 4, we illustrate the results of our experiments comparing the author productivity concentration of the five domains.

2. Concentration in Author Productivity Distributions

2.1 The Concept of Concentration and Inequality

Many measures of concentration have been proposed (e.g., Herfindahl, 1950; Theil, 1967; Atkinson,1970; Ray & Singer, 1973). However, the concept of concentration has not been clearly defined. Egghe and Rousseau (1991) have pointed out that proposed measures are thrown into a typical circular reasoning: "The notion of concentration is defined through the value of a measure used to measure concentration." Yoshikane (2000) redefined two types of concentration, namely relative concentration and absolute concentration. The former implies a relative inequality in distributions, while the latter refers to an absolute scale of distributions. In this study, by evaluating "relative" concentration (inequality) of productivity〔1〕, we aim to examine the variety of researchers' activity in the domain.

Egghe and Rousseau (1991) proposed some principles that any concentration measure must satisfy. Strictly speaking, since the number of events, which measures the absolute scale of distributions, is not taken into account, these principles apply to relative concentration rather than general concentration (Yoshikane, 2000). The proposed principles are as follows:

(1) When every event (e.g., author, in the case of author productivity data) appears with equal frequency, a concentration measure must have the value 0.
(2) For every permutation of the labels of events, the value of a concentration measure must be invariant.
(3) Even if the frequency is multiplied by $h(>0)$ for every event, the value of a concentration measure must be invariant.
(4) When we subtract $h(>0)$ from the frequency of a more frequent event and add h to that of a less frequent event, the value of a concentration measure must decrease.
(5) When we add $h(>0)$ to the frequency of the most frequent event, the value of a concentration measure must increase.
(6) If we add $h(>0)$ to the frequency of every event (where not all events are equal in frequency), the value of a concentration measure must decrease.

A concentration measure satisfying these principles has the attribute that if one takes from the poorer (less frequent events) and gives to the richer (more frequent events), the value of the measure increases. That is to say, the measure is appropriate for evaluating inequality, and hence, suits our purpose.

2.2 Concentration Measures

In this study, we use Pratt's measure (C) and the coefficient of variation (C_V) to evaluate inequality in the whole productivity distribution and inequality among core authors in the domain, respectively.

Pratt's measure:

$$C = \frac{2\left(\frac{V+1}{2} - q\right)}{V-1}$$

where

$$q = \sum_{i=1}^{V} i p_i$$

The coefficient of variation:

$$C_V = \frac{\sigma}{\mu}$$

In these formulae, V represents the number of authors (events), p_i represents the sample relative frequency of an author a_i (p_1, p_2, ..., p_V are ordered decreasingly), σ represents the standard deviation, and μ represents the mean frequency. Since they satisfy the following criteria, C and C_V are desirable as measures of relative concentration.

(i) Both measures satisfy all of the above-mentioned principles for "relative" concentration (Egghe & Rousseau, 1991).

(ii) Both measures are insensitive to "absolute" concentration, i.e., the number of authors (Yoshikane, 2000).

Thus, using these two measures, we are able to evaluate pure inequality in author productivity distributions.

2.3 Characteristics of Concentration Measures

The sensitivity of C_V is different from that of C, though both measures are sensitive to relative concentration. Specifically, C_V has a high sensitivity to the most frequent events (productive authors), while C is sensitive to all events (authors) equally. Using C and C_V, this study examines the tendency of the whole domain and that of the most productive authors in the domain. Below, we illustrate the characteristics of the measures by a simple simulation which is based on the geometric expression of the Lorenz curve.

The Lorenz curve, which visually expresses the degree of inequality, is the transition of the cumulative relative frequency where the frequencies of events (authors) are ordered increasingly. Two examples of the Lorenz curve are shown in Figure 1 (lines B-A1-C and B-A2-C). For convenience, this simulation uses polygonal lines instead of curves. Roughly speaking, the degree of inequality is expressed by the area enclosed by the Lorenz curve and the diagonal line (triangle B-A1-C or triangle B-A2-C) in the figure. When a distribution has high inequality of frequencies, the deflection from the diagonal line (line B-C) becomes large; that is, the area enclosed by the curve (the polygonal line) becomes large. The area corresponds to Pratt's measure C as well as Gini's index, a well-known measure in economics (Carpenter, 1979).

Triangles B-A1-C and B-A2-C are equal in area, because line D (0.4, 0.0) - E (1.0, 0.6) is parallel to line B (0.0, 0.0) - C (1.0, 1.0). Therefore, the value of C is invariant, as long as vertex A is on line D-E 〔2〕. Now, assume that vertex A moves from D to E along line D-E. Regarding Pratt's measure C, as mentioned above, the value is (approximately) invariant as long

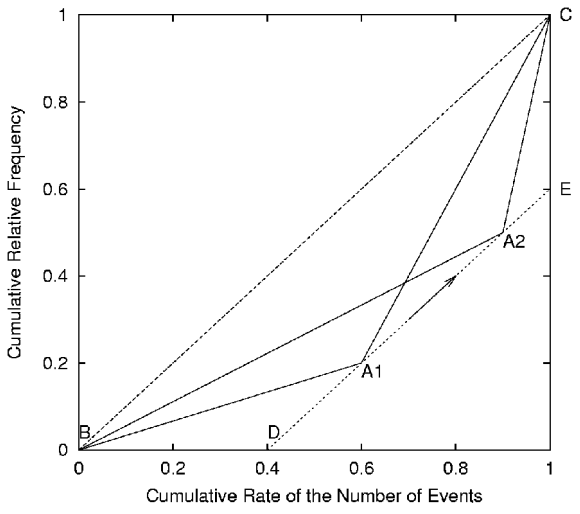

FIG. 1. Triangles representing the Lorenz curve.

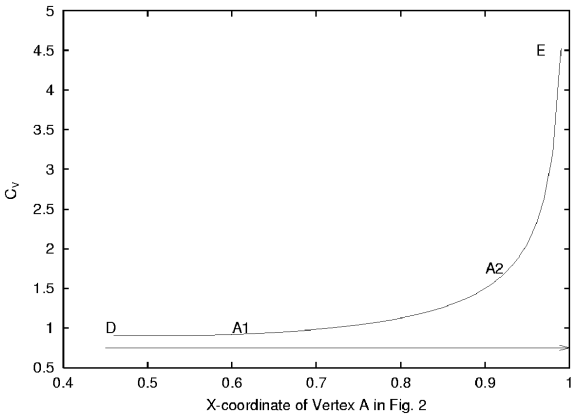

FIG. 2. Changes in the value of the coefficient of variation C_V.

as vertex A lies on line D-E. On the other hand, the coefficient of variation C_V systematically changes according to the location of vertex A. Figure 2 shows the change in the value of C_V in this simulation〔3〕. From this figure, it is observed that a distribution whose vertex A is located near E has an extremely high value of C_V. Vertex A is the point where the slope of the Lorenz curve changes. In other words, when the gap in terms of frequencies exists among the most frequent events〔4〕, the value of C_V becomes extremely high. Recalling that C is invariant in this simulation, we can confirm the high sensitivity of C_V to the most frequent events (the most productive authors).

3. Data

3.1 The Sample Size Dependency of Statistical Measures

Before entering into the explanation of data used in this study, we discuss the peculiar nature of bibliometric data to be considered while comparing different samples of different sizes, which causes the sample size dependency of statistical measures.

While enumerating their six principles for concentration measures, Egghe and Rousseau assumed that the number of events is constant. However, in the comparative analysis based on actual bibliometric data, including author productivity data, it is rather rare that all samples are equal in terms of the number of events. Even if samples are from the same population, the number of events is expected to change according to changes in the sample size. The sample size dependency is due to the fact that there are many low-frequency events (authors), and that not all the events (authors) in the population appear in the sample. Data of this type is called an LNRE (Large Number of Rare Events) sample (Khumarade, 1987).

Not only the number of events but also other measures calculated using frequencies of events crucially depend on the sample size, so long as there are unseen events that do not appear in the sample (Good, 1953; Good & Toulmin, 1956; Baayen, 2001). The coefficient of loss C_L is a convenient measure by which the reliability of data, as a sample, is checked. C_L gives the ratio of loss of the estimated number of events calculated using the sample relative frequencies as the estimates of population probabilities against the empirical number of events in a sample (Chitashvili & Baayen, 1993). Assuming the binomial model, C_L is calculated by:

$$C_L = \frac{V(N) - \hat{E}[V(N)]}{V(N)} = \frac{\sum_{m \geq 1} V(m,N)\left(1 - p\left(i_{[f(i,N)=m]}, N\right)\right)^N}{V(N)}$$

where

$f(i,N)$: frequency of an event (author) e_i in a sample of size N.

$p(i,N)$ $(=f(i,N)/N)$: sample relative frequency.

$V(m,N)$: the number of events appearing m times (authors who have each published just m papers).

When the coefficient of loss C_L is large, most statistical measures, including Pratt's measure C and the coefficient of variation C_V, systematically change according to changes in the sample size (Tweedie & Baayen, 1998). It is a typical peculiarity of LNRE samples that C_L is large. As for bibliometric data, for instance, Kageura (1998) have shown that the author productivity data has a large C_L.

Thus, in applying sample size dependent measures to author productivity data, we should carefully restrict the scope of interpretation to the "population" that a given "sample" constitutes by itself. We must not generalize the interpretation beyond the restricted population. If we are to describe the characteristics of what is beyond a given sample (e.g., the characteristics

of a whole domain, not that of data sampled from it), we need a statistical framework within which the dynamics of measures can be properly considered.

3.2 Samples

We extracted samples from a bibliographic database, the NACSIS database of academic conference papers, which had been provided by the National Institute of Informatics, Japan [5]. The records of conferences hosted from 1992 to 1997 by the following academic societies were extracted: Information Processing Society of Japan; Institute of Electrical Engineers of Japan; Architectural Institute of Japan; Society of Polymer Science, Japan; and Japan Society for Bioscience, Biotechnology, and Agrochemistry. We regard these as the data sampled from the entire data in each of the five domains (i.e., information processing, electrical engineering, architecture, polymer science, and biochemistry), and use these "samples" to compare the five domains in terms of author productivity concentration. The author-paper relation in the data is regarded as an indication of author productivity.

While dealing with the author-paper relation, we cannot avoid the issue associated with multiple authorship. This study credits each collaborating author with a full contribution, and it regards the total frequency of authors as the number of papers [6]. This is because our aim is not to evaluate researchers' contributions accurately [7], but to compare domains from the viewpoint of the amount of each researcher's activity, which is represented by the occurrence of names in the conference papers. So, in the following statistical arguments, the sample size refers to the total frequency of authors in a sample.

Table 1 shows the basic quantities of each sample, i.e., the total frequency of authors (N_0), the number of authors in a sample of size N_0 ($V(N_0)$), and the coefficient of loss (C_L).

TABLE 1. Comparison of the original samples.

	N_0	$V(N_0)$	C_L	$C(N_0)$	$C_V(N_0)$
Information Processing	79372	24271	0.225	0.562	2.112
Electrical Engineering	75685	25230	0.241	0.557	2.196
Architecture	166941	27143	0.159	0.640	2.065
Polymer Science	76413	16812	0.213	0.659	2.764
Biochemistry	71974	21315	0.229	0.582	2.070

C_L is shown in order to illustrate that our samples are statistically insufficient. In all the domains, except architecture, C_L exceeds 0.2. This means that the number of authors is underestimated by more than 20% if the population probabilities are estimated by the sample relative frequencies. In the data on architecture, C_L is smaller but far from negligible.

Besides, Table 1 gives Pratt's measure ($C(N_0)$) and the coefficient of variation ($C_V(N_0)$) at the original sample size N_0 for each domain. Among the five domains, polymer science has the highest value for both $C(N_0)$ and $C_V(N_0)$. That is, in polymer science, not only inequality between core authors and peripheral authors but also inequality among core authors is large. As for the remaining domains, i.e., information processing, electrical engineering, architecture, and biochemistry, a negative correlation between the two measures is observed.

Like this it is possible to compare the "samples" on the basis of the values shown in Table 1. However, we must recall that most measures, including $C(N)$ and $C_V(N)$, depend on the sample size N when C_L is large. That is to say, these results are no more than of a comparison of the original samples themselves. Therefore, we cannot generalize the characteristics of the whole of each domain from the results of a comparison based on limited samples (i.e., the data of the papers from 1992 to 1997 in each academic conference).

4. Analysis

4.1 Methodology

In the previous section, we have shown that our samples are belong to an LNRE sample, where statistical measures change systematically, not randomly, according to changes in the sample size. Therefore, we cannot directly compare the characteristics of the populations on the basis of the values of measures that are calculated from the samples themselves.

There are two approaches that can be used for comparative analysis on the basis of LNRE samples.

The first is to adopt sample size invariant measures. For example, Yule's characteristic constant K (Yule, 1944) is well-known as a sample size invariant measure (Yoshikane & Kageura, 1999). The second is to normalize the sample size by taking random sub-samples of the larger samples, or tracing the "developmental profiles" of measures and examining their features in their totality (Yoshikane, Kageura, & Tsuji, 2003). The former approach is useful, but problematic in that there are very few measures that do not depend on the sample size. At least as for concentration in the whole productivity distribution, we have been unable to find a proper sample size invariant measure, which is sensitive to all events (authors) equally. As mentioned above, Pratt's measure C as well as the coefficient of variation C_V depends on the sample size.

Thus, in this study, we take the latter approach. We carried out 1,000 random Monte-Carlo sub-samplings of the original samples for 20 equally-spaced intervals in order to observe the change in each measure. We aim to describe the characteristics of the domain itself beyond the given sample size by observing the developmental profiles of the two measures, $C(N)$ and $C_V(N)$.

4.2 Results

Concentration in the Whole Productivity Distribution

Figure 3 plots the developmental profiles of Pratt's measure $C(N)$ obtained by the random Monte-Carlo sub-samplings. In all the domains, it is observed that $C(N)$ increases systematically according to growth in the sample size N [8]. The figure implies that when we compare different samples of different sizes, the result, unfortunately, varies with the sample size of each domain. For example, comparing architecture and biochemistry at the same sample size, the former has lower values than the latter. Contrary to this, while comparing their original samples on the basis of $C(N_0)$, architecture is placed higher than biochemistry (see Table 1). It should be noted that this is due to the original sample size of architecture, which is twice as large as that of biochemistry.

Upon normalizing the sample size, we clearly observe that polymer science has the highest values of $C(N)$ among the five domains. That is to say, inequality between core authors and peripheral authors is large in this domain. Polymer science is followed by biochemistry and architecture. Information processing and electrical engineering have the lowest values of $C(N)$. Judging from the correspondence between information processing and electrical engineering in Figure 3, we can say that with regard to concentration in the whole productivity distribution, the two domains are almost equal.

Behavior of Core Authors in the Domain

The coefficient of variation $C_V(N)$ is plotted in Figure 4 in the same way that $C(N)$ is plotted in Figure 3. It is observed that $C_V(N)$ as well as $C(N)$ changes according to changes in the sample size N.

Tracing the developmental profiles in Figure 3, we

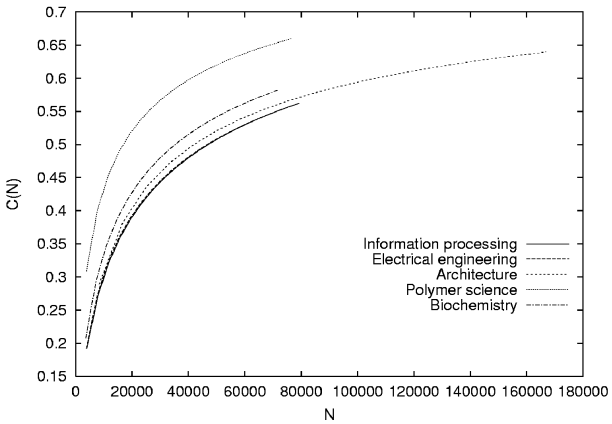

FIG. 3. Change of the values of Pratt's measure $C(N)$.

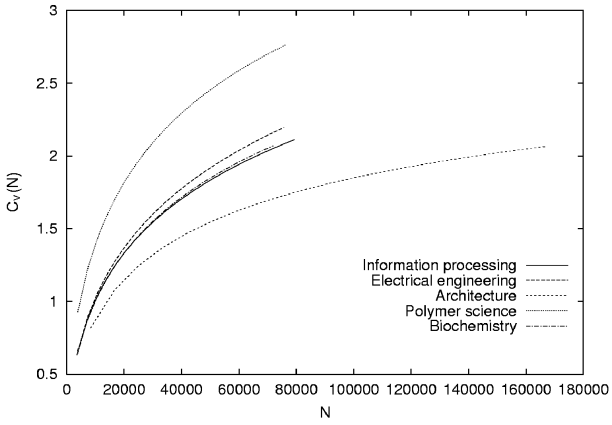

FIG. 4. Change of the values of the coefficient of variation $C_V(N)$.

can observe that, whatever sample size we take, polymer science is much higher than the other domains for the coefficient of variation $C_V(N)$ just as it is for Pratt's measure $C(N)$. That is, not only the inequality between core authors and peripheral authors but also that among core authors is large in this domain. Polymer science is followed by electrical engineering, biochemistry, and information processing. Architecture has the lowest value of $C_V(N)$ whatever sample size we take. In the comparison of the five domains on the basis of $C_V(N)$ at the same sample size, the degree of concentration exhibits the following descending order: polymer science, electrical engineering, biochemistry, information processing, and architecture. Biochemistry and information processing are relatively close to each other.

Overall Observations

In order to visualize the characteristics of the five domains more clearly, we plotted the pattern of domain characteristics at the same sample size in Figure 5. Figure 5 shows four snapshots of the pattern at the sample size N = 10,000, 30,000, 50,000, 70,000. In Figure 5, at any sample size, common features are observed, i.e., polymer science exceeds the other domains both in Pratt's measure $C(N)$ and in the coefficient of variation $C_V(N)$, information processing and electrical engineering have the lowest values of $C(N)$, architecture has the lowest values of $C_V(N)$, and biochemistry is located at the center of the five domains.

It is interesting to see that a positive correlation between the two concentration measures is not necessarily observed. For instance, with regard to $C_V(N)$, architecture has the lowest values among the five domains. As for $C(N)$, on the other hand, architecture ranks higher than information processing and electrical engineering.

Recalling the characteristics of $C(N)$, we can assume that the three domains located lowest in Figure 5 (i.e., information processing, electrical engineering, and architecture) are similar with respect to the "mass" behavior of authors, i.e., in these domains their

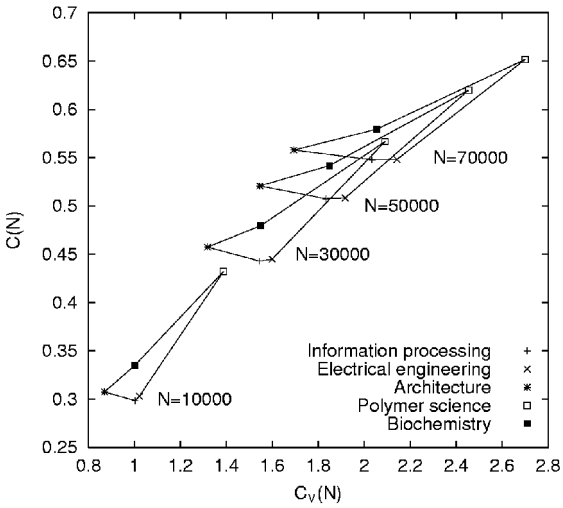

FIG. 5. Comparison of the five domains by $C(N)$ and $C_V(N)$.

productivity patterns are more homogeneous. Regarding architecture, which has the lowest values of $C_V(N)$, the productivity patterns of core authors are also homogeneous. Contrary to architecture, polymer science, which is located in the upper right-hand corner of Figure 5, is a domain where both the mass behavior of authors, including peripheral ones, and the behavior of core authors are heterogeneous.

5. Conclusions

In this article, we compared the author productivity concentration of the five domains, taking into consideration the sample size dependency of measures. By observing the transitions of Pratt's measure and the coefficient of variation according to changes in the sample size, we have analyzed (1) concentration in the whole productivity distribution and (2) behavior of core authors in the domain.

The characteristics of each domain shown in our analysis can be summarized as follows. Polymer science shows higher concentration both in mass behavior and in the behavior of core authors. Information processing and electrical engineering are more homogeneous in terms of mass behavior, while architecture is more homogeneous in terms of the behavior of core authors. Biochemistry ranks at an intermediate level in terms of author productivity concentration. The behavior of core authors is not necessarily correlated with the "mass" behavior of

authors, including peripheral ones. From this result, we can confirm the independency of core researchers in the domain.

Note

〔1〕 We use "relative concentration" and "inequality" interchangeably.

〔2〕 Pratt's measure *C* is approximately, but not completely, equal to Gini's index, which is calculated on the basis of the area enclosed by the Lorenz curve. Therefore, to be precise, *C* is not completely, but approximately invariant on this condition.

〔3〕 This simulation is carried out under the condition that the number of events is 100 and that the total frequency is 10,000.

〔4〕 As the frequencies of events are ordered "increasingly" in the Lorenz curve, the most frequent events appear near E.

〔5〕 After 2003, it is provided by the Japan Science and Technology Agency.

〔6〕 Incidentally, Lotka (1926) as well as Pao (1986) uses the first author only.

〔7〕 For the purpose of evaluating their contributions accurately, it may be better to use an "adjusted count" (Lindsey, 1980; 1982), which credits each collaborating author with some weight according to the number of co-authors. However, an adjusted count produces decimal fractions in the count data and makes the manipulation of the data technically and conceptually much more difficult.

〔8〕 According to Pratt (1977), *C* is independent of both the number of sources (e.g., authors) and that of items (e.g., papers). Even though *C* is a measure that is normalized between 0 and 1 without regard to the number of papers, it actually depends on the number of papers, i.e., the sample size, as shown here.

References

Atkinson, A. B. (1970). On the measure of inequality. *Journal of economic theory*, 2(3), 244-263.

Baayen, R. H. (2001). Word frequency distributions. Dordrecht: Kluwer Academic Publishers.

Carpenter, M. P. (1979). Similarity of Pratt's measure of class concentration to the Gini index. *Journal of the American Society for Information Science*, 30(2), 108-110.

Chitashvili, R. J., & Baayen, R. H. (1993). Word frequency distributions. In: L. Hrebicek, & G. Altmann (Eds.), Quantitative text analysis (pp. 54-135). Trier: Wissenschaftlicher Verlag.

Egghe, L. (1987a). The relative concentration of a journal with respect to a subject and the use of online services in calculating it. *Journal of the American Society for Information Science*, 38(4), 288-297.

Egghe, L. (1987b). Concentration places, concentration evolutions, and online information retrieval techniques for calculating them. *Information Processing & Management*, 24(2), 109-121.

Egghe, L., & Rousseau, R. (1991). Transfer principles and a classification of concentration measures. *Journal of the American Society for Information Science*, 42(7), 479-489.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4), 237-264.

Good, I. J., & Toulmin, G. H. (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika*, 43(1), 45-63.

Herfindahl, O. C. (1950). Concentration in the Steel Industry, PhD thesis, Columbia University, New York.

Kageura, K. (1998). Some characteristics of bibliometric samples: An examination of Lotka-type data. *Annals of the Society of Library Science*, 44(3), 97-110.

Khumarade, E. V. (1987). The statistical analysis of Large Number of Rare Events. Report MS-R8804, Department of Mathematical Statistics. Amsterdam: Center for Mathematics and Computer Science.

Kishida, K. (1996). Some characteristics of scientometric indicators. *Journal of Japan Indexers Association*, 20(2), 1-11.

Lindsey, D. (1980). Production and citation measures in the sociology of science: The problem of multiple authorship. *Social Studies of Science*, 10(2), 145-162.

Lindsey, D. (1982). Further evidence for adjusting for multiple authorship. *Scientometrics*, 4(5), 389-395.

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, 16(12), 317-323.

Pao, M. L. (1986). An empirical examination of Lotka's law. *Journal of the American Society for Information Science*, 37(1), 26-33.

Pratt, A. D. (1977). A measure of class concentration in bibliometrics. *Journal of the American Society for Information Science*, 28(5), 285-292.

Ray, J. L., & Singer, J. D. (1973). Measuring the concentration of power in the international system. *Sociological Methods & Research*, 1(4), 403-437.

Theil, H. (1967). Economic and information theory. Amsterdam: North-Holland Publishing Company.

Tweedie, F. J., & Baayen, R. H. (1998). How variable may a constant be?: Measures of lexical richness in perspective. *Computers and the Humanities*, 32, 323-352.

Yoshikane, F., & Kageura, K. (1999). On the potential of sample size invariant measures for the comparative analysis of author productivity data. In C. A. Macias-Chapula (Ed.), Proceedings of the 7th Conference of the International

Society for Scientometrics and Informetrics (ISSI ' 99) (pp. 547-557). Colima: Universidad de Colima.

Yoshikane, F. (2000). Concentration in bibliometric distributions: The notion of concentration and concentration measures. *Journal of Japan Society of Library and Information Science*, 46(1), 18-32.

Yoshikane, F., Kageura, K., & Tsuji, K. (2003). A method for the comparative analysis of concentration of author productivity, giving consideration to the effect of sample size dependency of statistical measures. *Journal of the American Society for Information Science and Technology*, 54(6), 521-528.

Yule, G. U. (1944). The statistical study of literary vocabulary. Cambridge: Cambridge University Press.