Essays of an Information Scientist, Vol:4, p.476-483, 1979-80 Current Contents, #19, p.5-12, May 12, 1980

Current Comments

Bradford's Law and Related Statistical Patterns

Number 19 May 12, 1980

From time to time, I refer to "Bradford's law." Those connected with information science usually recognize this concept. Actually called Bradford's *law of scattering*, it describes how the literature on a particular subject is scattered or distributed in the journals.[1]

Bradford's law is one of several statistical expressions which try to describe the workings of science by mathematical means. While each "law" applies to a different specific phenomenon, they all tend to demonstrate one thing—that a few (journals, scientists, etc.) account for the many (articles, citations, etc.). In practical terms, this means that there are diminishing returns in trying to do anything exhaustively. For example, in information retrieval we know that neither a library nor a service like *Current Contents® (CC®)* can cost-effectively cover every journal that might be relevant to a given area.[2]

One of the statistical tools related to Bradford's law is the Pratt index. This index purports to measure the degree to which papers on a given subject are concentrated within a journal collection.[3] Then there's Lotka's law, which describes the productivity distribution among scientists, as, for example, how many scientists will author ten or more papers.[4]

These measures are important for anyone interested in the sociology of science, but they are also similar to many concepts found in other fields. For example, the frequency distribution predicted by Bradford's law so closely resembles Zipf's law of word frequency[5] that it is often referred to as the Bradford-Zipf distribution.[6] The Pratt index is virtually identical to the Gini index,[7] which is known to economists as the measure of wealth concentration among members of a society.[8] And the formula for Lotka's law resembles another economic paradigm—Pareto's law of income distribution.[9] (p. 49)

Several years ago, Derek Price proposed a unifying theory for all of these statistical "laws."[10] His paper, which appeared in the *Journal of the American Society for Information Science*, contained a formula which describes equally well such diverse behaviors as article scatter, frequency of word use, and even an article's citation accumulation. B.C. Brookes, University College, London, has also described "an empirical law of social behavior which pervades all social activities."[11] These unifying theories may eventually render it inappropriate to discuss the above laws individually. Nevertheless, much continues to be written about all of them.

Although we do not use these laws directly when determining our journal coverage, our awareness of their underlying principles has guided us in designing the optimum coverage for *CC*, *Science Citation Index® (SCI®)*, and all of ISI®'s services. Without this awareness, we would be guilty of perpetuating what Bradford described as "documentary chaos."[1]

In the discussion that follows, you will no doubt recognize statistical concepts well known in other fields. But the average reader, like myself, is not a statistician. So in describing these concepts, I won't dwell on the mathematics involved. Rather, I want to discuss them in terms that will give the non-specialist an intuitive feeling for what these laws describe. Those interested in the mathematics should turn to the appendix which follows this essay.

Bradford's law is perhaps the best known of all the bibliometric concepts discussed here. A huge body of literature has been written on the subject.[12-20] It is most valuable to librarians who are faced with the cost-benefit considerations of additional journal coverage.[12]

Samuel C. Bradford first formulated his law in 1934.[21] But it did not receive wide attention until the first publication of his book *Documentation* in 1948.[1] The law derives its universality from the basic unity of science—that is, that every scientific field is related, however remotely, to every other field. If you want to compile a bibliography on any subject, you will find that there is always a small group of core journals that account for a substantial percentage (1/3) of the articles on that subject or discipline. Then there is a second larger group of journals that account for another third while a much larger group of journals picks up the last third.

A physical analogy of the situation described by Bradford would be a comet, with the nucleus representing the core journals of a literature and the debris and gas molecules of the tail representing additional journals that sometimes publish material relevant to the subject. As subjects or disciplines get larger the pursuit of complete coverage takes you farther and farther afield. Thus, to cover all of chemistry, *Chemical Abstracts* claims to scan over 10,000 journals, though it is well known that a large percentage of its coverage is derived from 1,000.

In *Documentation*, Bradford analyzed a four-year bibliography of references to articles in applied geophysics. He listed the journals containing references to that field in descending order of productivity. (See Table 1.) He then divided the list into three "zones," each containing roughly the same number of references. Bradford observed that the number of journals contributing references to each zone increased by a multiple of about five. Specifically, the first zone contained nine journals which contributed 429 references. The second contained 59 journals producing 499 references. In the third zone 258 journals provided 404 references. Bradford found a similar pattern of reference scatter in the field of lubrication. On the basis of these observations, Bradford wrote..."the numbers of periodicals in the nucleus and succeeding zones will be as 1, n, n^2...."[1] (p. 116) For applied geophysics then, the number of journals in each zone was roughly proportionate to 1, 5, 25.

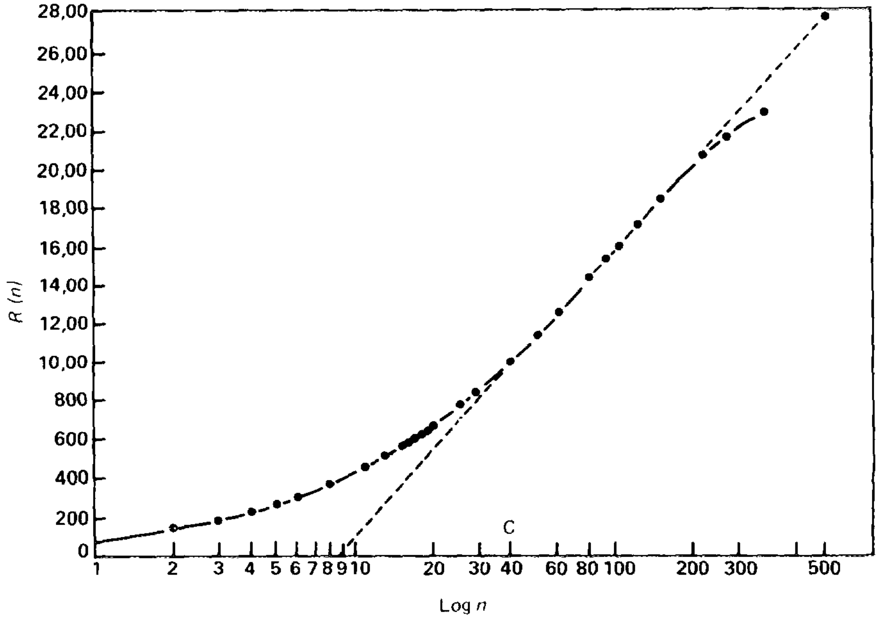
The graph in Figure 1 illustrates Bradford's law for articles on tropical and subtropical agriculture found in *Tropical Abstracts* during 1970. The graph is taken from a paper by S.M. Lawani of

Table 1: S.C. Bradford's applied geophysics data. **A** = number of journals producing the corresponding articles in column B. **B** = number of relevant articles found in each journal. **C** = journal rank in descending order of reference productivity. **D** = running total of references found in all journals.

A	B	C	D
1	93	1	93
1	86	2	179
1	56	3	235
1	48	4	283
1	46	5	329
1	35	6	364
1	28	7	392
1	20	8	412
1	17	9	429
4	16	13	493
1	15	14	508
5	14	19	578
1	12	20	590
2	11	22	612
5	10	27	662
3	9	30	689
8	8	38	753
7	7	45	802
11	6	56	868
12	5	68	928
17	4	85	996
23	3	108	1,065
49	2	157	1,163
169	1	326	1,332

Reprinted by permission from: **Bradford S C.** *Documentation.* Washington, DC: Public Affairs Press, 1950. Copyright 1950, by Public Affairs Press.

Figure 1: The distribution of tropical and subtropical agriculture articles in 1970. Journals containing relevant articles are plotted logarithmically along the horizontal axis. The running sums of articles are plotted along the vertical axis.



Reprinted by permission from: **Lawani S M.** Bradford's law and the literature of agriculture. *Int. Libr. Rev.* 5:341-50, 1973. Copyright by Academic Press, Inc. (London) Ltd.

the International Institute of Tropical Agriculture, Nigeria.[13]

The entire collection in Figure 1 consists of 2,284 articles found in 374 journals. Notice that the line initially appears as an upward curve before it becomes linear. This is typical of Bradford graphs. The area described by the curving line is usually regarded as the "nuclear zone," or journal core. The number of journals found in the nuclear zone varies according to the bibliography. In this case, 40 journals containing 1,025 articles are in the nuclear zone. The number of articles retrieved from each additional journal steadily decreases. For example, 175 journals contribute 1,962 articles—about 11 each. The next 25 journals contribute only 75 additional articles, or 3 articles apiece.

Bibliographic studies in a number of fields confirm that the dispersion of articles throughout a set of journal titles conforms to a statistical distribution of the type described by Bradford.[13-16] But arriving at a precise mathematical expression for this distribution proved to be a matter of controversy.

In 1967, Ferdinand F. Leimkuhler, Purdue University, formally derived the equation for plotting a Bradford curve.[17] The formula he provided was the first general expression of article scatter. But the next year, Brookes complained that the formula required too much tedious computation to be of practical use to librarians. "In fact," wrote Brookes, "it was the exasperation evoked by an attempted practical application of Leimkuhler's formulae" that led him to seek a simpler formulation of the Bradford distribution.[6] Both Leimkuhler's equation and Brookes' simplified formula appear in the appendix.

It was Elizabeth A. Wilkinson, University College, London, who, in 1972, suggested that the formulas provided by Leimkuhler and Brookes did not really describe the same phenomenon.[18] The discrepancy arose from an error that Bradford himself had made in *Documentation*. Bradford provided both a *graphical* representation of his applied geophysics data, and the *verbal* expression of his law previously quoted in this essay. In 1948, B.C. Vickery, University College, London, was the first to notice that the graph provided by Bradford was not necessarily equivalent to the statement "...the numbers of periodicals in the nucleus and succeeding zones will be as 1, n, n^2...."[19] Now Wilkinson observed that Leimkuhler had derived his distribution function from Bradford's verbal expression, while Brookes derived his formula from Bradford's graphical representation. In comparative tests, she found that the graphic formulation more closely conformed to empirical data than the verbal expression in six of eight cases.[18]

When Brookes published his simplified derivation of the Bradford distribution, he recognized its similarity to another empirical law, the one named for George K. Zipf. At about the same time that Bradford wrote *Documentation*, Zipf published his book *Human Behavior and the Principle of Least Effort*.[5] Among the topics treated in the book was the frequency with which words occur in a given piece of literature. Zipf arranged the 29,899 different words found in Joyce's *Ulysses* in descending order of their frequency of occurrence. (See Table 2.) Then to each word he assigned a rank, from r=1 (most frequently occurring word) to r=29,899 (least frequently occurring). He found that by multiplying the numerical value of each rank r by its corresponding frequency f, he obtained a product, C, which was constant throughout the entire list of words. The formula for Zipf's law is thus rf=C.

Table 2: The distribution of words in Joyce's *Ulysses* illustrating Zipf's law.

I Rank (*r*)	II Frequency (*f*)	III Product of I and II ($r \times f = C$)
10	2,653	26,530
20	1,311	26,220
30	926	27,780
40	717	28,680
50	556	27,800
100	265	26,500
200	133	26,600
300	84	25,200
400	62	24,800
500	50	25,000
1,000	26	26,000
2,000	12	24,000
3,000	8	24,000
4,000	6	24,000
5,000	5	25,000
10,000	2	20,000
20,000	1	20,000
29,899	1	29,899

Reprinted from: **Zipf G K.** *Human behavior and the principle of least effort.* New York: Hafner, 1972.

Zipf's law does not describe a Bradford distribution precisely. It is true that multiplying a journal's rank by the number of articles it contributes will yield a rough constant. But this applies only for the straight line portion of a Bradford curve. The nuclear zone does not conform to Zipf's rf=C. Brookes suggests a reason for this deviation from Zipf's law. He notes that journals "lack certain elements of statistical uniformity."[6] For example, journals are not all issued at equal intervals. Nor do they contain the same number of articles within their different issues. Thus, Bradford's law is reliable only for large collections. The few journals found in the nuclear zone are apt to deviate from the law.[6]

The non-conformity of the nuclear zone is not the only anomaly associated with Bradford's law. In Figure 1 notice that the straight line begins to droop at about the 250th journal. This is called the Groos droop, named for its discoverer, Ole V. Groos.[20] The droop consistently appears among many different sets of collected data. Several reasons have been advanced to explain it. The one most commonly accepted is that the droop reflects the necessary incompleteness of the bibliography.[13] Theoretically, the search could be extended until the droop is restored to linearity. But another theory held by those "who have painstakingly searched the literature, and other realists,"[6] states that the Bradford law is correct only as far as it goes, and that the droop is an integral part of article scatter.[6] In any case, the droop occurs too far beyond the point of diminishing returns to be of any consequence to the practical librarian considering additional journal coverage.[22]

Brookes' expression of the Bradford distribution has gained wide acceptance. But it does not predict *a priori* what the exact results of a literature search will be. His formula contains a variable which describes the slope of a Bradford curve. This variable is constant within a journal collection. But its numerical value differs from bibliography to bibliography. Moreover, extending or shortening the time span of the bibliography can also affect the constant's numerical value. In other words, one cannot determine how many relevant articles will be found in a given journal until a bibliography has actually been compiled. This shortcoming of Bradford's law as a numerical predictor has prompted M. Carl Drott, Drexel University, Philadelphia, to call it "both a pervasive and an elusive phenomenon."[15] More to the point, Brookes himself recently wrote, "We may...be mistaken in continuing to search for that single formulation embracing all Bradford phenomena which has eluded capture for more than forty years."[23]

Nevertheless, Bradford's law provides us with a clear description of a basic bibliometric behavior. In so doing, it makes some sense out of the "documentary chaos." The forces underlying Bradford's law allow one to make other observations. For example, one can look at bibliometric *concentration*, as opposed to scatter. This is the approach I took in formulating Garfield's "law of concentration."[24,25] This law, more properly called an axiom, simply points out that for any field of science, articles are concentrated essentially within the same highly-cited or multidisciplinary journals.

Specifically, I wrote that the "comet's tail" of one discipline's literature consists, in large part, of the cores of the literature of other disciplines. So large is the overlap between disciplines, in fact, that the core literature for all scientific disciplines involves a group of no more than 1,000 journals, and may involve as few as 500. This means that a good general science library that covers the core literature of all disciplines need not have any more journals than a good special library that covers all the literature of a single discipline.[25]

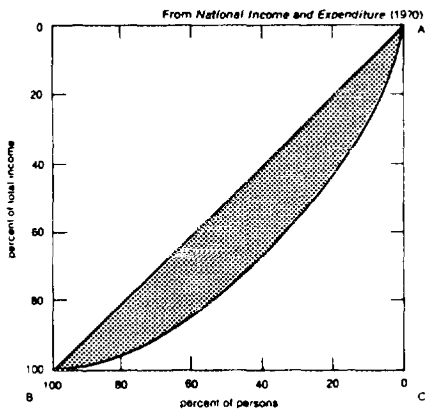
This also means that an index may cover only a limited number of journals, as long as they are the core journals, and still provide excellent coverage. In fact, we have applied this principle to our coverage in the *SCI*. The *SCI* covers more than 3,000 journals, far more than the 1,000 core journals but far less than the total number of journals published, which is at least 10,000. Yet we can still be sure that we are providing access to 90% of the significant journal literature.

We have noted that the precise slope of a Bradford curve will differ among document collections. Thus, 10% of a journal collection might contribute 50% of the relevant articles in one field, while in another field, 20% of a journal collection might contribute the same percentage of articles. Is it possible to measure the degree to which articles on a particular subject are concentrated within the literature? An index recently proposed by Allan Pratt, Indiana University, provides such a measure. Although Pratt addressed the problem of how journal articles in a broad subject field are concentrated within its constituent sub-specialties, he notes that there are other concentration phenomena that can be measured with the index.[3] With a simple substitution of the terms used, the index can be applied to measure Bradford type article scatter.

The Pratt index is basically a ratio between two mathematical expressions. The first expression represents a *hypothetical* case in which articles in a collection are distributed equally among a field's sub-specialties. The second expression describes how the articles are actually distributed among those subspecialties. The degree of literature concentration is expressed as a fraction between zero and one. The former number represents equal distribution, while the latter represents total concentration.[3]

Soon after Pratt published his index, Mark Carpenter, Computer Horizons, New Jersey, noticed its similarity to the Gini index.[7] The Gini index is named after the Italian economist, Corrado Gini, who first formulated it in 1908. It is based upon the construction of a Lorenz curve, which shows how a society's wealth is distributed among its people.[8] Figure 2 shows a Lorenz curve

Figure 2: Lorenz curve of personal income distribution, Great Britain, 1967. Straight line represents hypothetical equality. Curved line represents actual income distribution.



Reprinted by permission from: **Champernowne D G.** Distribution of wealth and income. *Encyclopaedia Britannica.* Chicago: H.H. Benton, Publ., 1974, Vol. 19, p. 677, and Central Statistical Office, London.

which represents the distribution of income among the population of Great Britain in 1967. The straight line on the graph represents a hypothetical case in which everyone has an equal amount of income. The curved line represents the actual distribution of wealth. To compute the Gini index, one calculates the ratio between the shaded area and the entire triangle described by the "ideal" equal distribution line. Like the Pratt index, the Gini is set between zero and one to represent respectively total equality and total concentration of wealth.

But not everyone is convinced that a Gini-type index can provide a meaningful description of bibliometric concentration. Carl Drott, for example, observes that the results obtained by applying Pratt's formula depend more on sample size than on any intrinsic concentration of the literature.[26] Drott notes that the numerical value of Pratt's index for a subject literature spanning one year will differ from, say, a four-year collection of the same subject literature. In short, anything that affects the size of the literature collection will affect Pratt's measure of concentration for that collection.[26] A similar case has been made by J. Hustopecký and J. Vlachý in a paper which appeared in *Scientometrics*.[27]

Distribution frequencies of the Bradford-Zipf type are also evident in other statistical phenomena. Lotka's law, for example, describes the productivity of scientists within a given population.[4] Productivity is defined here as the number of papers a scientist publishes within a given time. In 1926, Alfred J. Lotka observed that the distribution of scientific authorship follows an inverse square formula. That is, the number of scientists who author n papers will be $1/n^2$ of those who author just one paper. Thus, for every 100 authors who produce just one paper, 25 will produce two, 11 will produce three, and so on. At least one study has shown that Lotka's law also applies to the humanities literature.[28]

In his classic book *Little Science, Big Science,* Price wrote that Lotka's simple inverse square formula tends to overestimate the number of high-productivity authors.[9] In fact, the number of people within the highest productivity range falls off more nearly by the inverse cube, rather than the inverse square. Price also observed similarities between Lotka's law and Pareto's law of income distribution, which states that cumulative figures for income follow a $1/n^{1.5}$ law.[9]

Several studies of the applicability of Lotka's law have produced negative results.[29-31] But in a recent paper, K. Subramanyam, Drexel University, ascribed these results to such causes as sample selection or the manner in which multiple-authored papers were treated.[32]

Subramanyam acknowledges, as does Price, that authorship alone is a poor measure of productivity. As Price put it, "Who dares to balance one paper of Einstein on relativity against even a hundred papers by John Doe, Ph.D., on the elastic content of the various timbers (one to a paper) of the forests of Lower Basutoland?"[9] (p. 40)

Bradford's, Zipf's, Lotka's, and Pareto's laws were independently formulated to explain disparate phenomena. It seems more than coincidence that they should closely resemble each other. One wonders if they are not all governed by a single underlying principle. As mentioned earlier, Price has taken a step toward identifying such a principle. He has proposed a unifying conceptual model with which to view such diverse phenomena as article scatter, wealth distribution, scientific authorship, and even an article's citation accumulation.[10]

Price describes his "theory of cumulative advantage processes" as having a single-edged Matthew effect, i.e., success is rewarded but failure has no consequences. Price provides a formula to describe a population of individuals trying to achieve a goal. In the case of Lotka's law, the individual is a scientist, and the goal is publication of a paper. In the case of Bradford's law, the individual is a journal, and the goal is the yielding of a relevant article.

Price's equation is presented in the appendix. It is "rigged" so that if an individual (scientist, journal) is successful (publishes, yields an article) on one attempt the probability of success on subsequent attempts increases. However, a failure does not diminish the probability of success on the next attempt, since non-publication or non-yielding of an article are undefinable events.[10]

The cumulative advantage process described by Price is incredibly pervasive. Statistical distributions of the success-breeds-success type apply to such unlikely phenomena as the number of fossils recovered from a sample of rock strata,[33] and the number of times that books in a library are used.[34]

Our discussion of Bradford's law and related statistical expressions must be limited. Nevertheless, I hope that it will be a useful summary to informationists and expose our mainly scientist-scholarly audience to the kinds of concerns that preoccupy people in the fields of bibliometrics and scientometrics—as well as producers of information services. Bradford's law shows us that our finite resources would be depleted in pursuit of total coverage, and that the users of our services would retrieve little more of value to them if we went to the expense of covering "everything." Since Bradford's law is useful to us, it is indirectly useful to you. But you don't have to understand Ohm's law to turn on a lamp or Bradford's law to use *Current Contents.*

* * * * *

My thanks to Linda Cooper and Thomas Di Julia for their help in the preparation of this essay. ©1980 ISI

Appendix of Equations

1. Leimkuhler's Bradford distribution: $F(x) = \dfrac{\ln(1+\beta x)}{\ln(1+\beta)}$

x denotes the fraction of documents in a collection which are most productive, $0 \leq x \leq 1$. F(x) denotes the proportion of total productivity contained in the fraction x. The parameter β is related to the subject field and the completeness of the collection.

2. Brookes' simplified Bradford-Zipf distribution: $R(n)=k \log n/s$

$R(n)$ denotes the number of relevant papers contributed by the journal ranked n. k is a constant which determines the slope of the Bradford curve and is related to the document collection. s is a constant which determines where the straight line, if extended, would intersect the horizontal axis.

3. Price's cumulative advantage distribution: $f(n)= (m+1)\beta (n, m+2)$

n = number of successes; f(n) = fraction of individuals with n successes; m = a constant for all individuals in a population for all n.

REFERENCES

1. **Bradford S C.** *Documentation.* Washington, DC: Public Affairs Press, 1950. 156 p.
2. **Garfield E.** How do we select journals for *Current Contents? Current Contents* (45):5-8, 5 November 1979.
3. **Pratt A D.** A measure of class concentration in bibliometrics. *J. Amer. Soc. Inform. Sci.* 28:285-92, 1977.
4. **Lotka J A.** The frequency distribution of scientific productivity. *J. Wash. Acad. Sci.* 16:217-23, 1926.
5. **Zipf G K.** *Human behavior and the principle of least effort.* New York: Hafner, 1972. 573 p.
6. **Brookes B C.** The derivation and application of the Bradford-Zipf distribution. *J. Doc.* 24:247-65, 1968.
7. **Carpenter M.** Similarity of Pratt's measure of class concentration to the Gini index. *J. Amer. Soc. Inform. Sci.* 30:108-10, 1979.
8. **Gini C.** On the measure of concentration with special reference to income and wealth. (Sisam C H, ed.) *Research conference on economics and statistics.* Colorado Springs: Colorado College, 1936. p. 73-80.
9. **Price D J D.** *Little science, big science.* New York: Columbia, 1963. 118 p.
10. --------------. A general theory of bibliometrics and other cumulative advantage processes. *J. Amer. Soc. Inform. Sci.* 27:292-306, 1976.
11. **Brookes B C.** Theory of the Bradford law. *J. Doc.* 33:180-209, 1977.
12. **Weinstock M.** Bradford's law. (Letter to the editor.) *Nature* 233:434, 1971.
13. **Lawani S M.** Bradford's law and the literature of agriculture. *Int. Libr. Rev.* 5:341-50, 1973.
14. **Alabi G.** Bradford's law and its application. *Int. Libr. Rev.* 11:151-8, 1979.
15. **Drott M C, Mancall J C & Griffith B C.** Bradford's law and libraries: present applications—potential promise. *Aslib Proc.* 31:296-304, 1979.
16. **Aiyepeku W O.** Bradford distribution theory—compounding of Bradford periodical literatures in geography. *J. Doc.* 33:210-19, 1977.
17. **Leimkuhler F F.** The Bradford distribution. *J. Doc.* 23:197-207, 1967.
18. **Wilkinson E A.** The ambiguity of Bradford's law. *J. Doc.* 28:122-30, 1972.
19. **Vickery B C.** Bradford's law of scattering. *J. Doc.* 4:198-203, 1948.
20. **Groos O V.** Bradford's law and the Keenan-Atherton data. *Amer. Doc.* 18:46, 1967.
21. **Bradford S C.** Sources of information on specific subjects. *Engineering* 137:85-6, 1934.
22. **Buckland M K.** *Book availability and the library user.* New York: Pergamon, 1975. 196 p.
23. **Brookes B C.** The Bradford law: a new calculus for the social sciences? *J. Amer. Soc. Inform. Sci.* 30:233-4, 1979.
24. **Garfield E.** The mystery of the transposed journal lists—wherein Bradford's law of scattering is generalized according to Garfield's law of concentration. *Current Contents* (31):5-6, 4 August 1971. (Reprinted in: **Garfield E.** *Essays of an information scientist.* Philadelphia: ISI Press, 1980. Vol. 1. p. 222-3.)
25. -------------. *Citation indexing: its theory and application in science, technology, and humanities.* New York: John Wiley & Sons, 1979. 274 p.
26. **Drott M C.** Telephone communication. 1 April 1980.
27. **Hustopecký J & Vlachý J.** Identifying a set of inequality measures for science studies. *Scientometrics* 1:85-98, 1978.
28. **Murphy L J.** Lotka's law in the humanities? *J. Amer. Soc. Inform. Sci.* 24:461-2, 1973.
29. **Radhakrishnan T & Kernizan R.** Lotka's law and computer science literature. *J. Amer. Soc. Inform. Sci.* 30:51-4, 1979.
30. **Voos H.** Lotka and information science. *J. Amer. Soc. Inform. Sci.* 25:270-2, 1974.
31. **Worthen D B.** Short-lived technical literatures: a bibliometric analysis. *Methods Inform. Med.* 17:190-8, 1978.
32. **Subramanyam K.** Lotka's law and the literature of computer science. *IEEE Trans. Prof. Commun.* 22:187-9, 1979.
33. **Brookes B C.** Reply to: Smith D A. (Letter to the editor.) *J. Doc.* 25:155, 1969.
34. **Bulick S.** Book use as a Bradford-Zipf phenomenon. *Coll. Res. Libr.* 39:215-9, 1978.