STATISTICAL ANALYSIS OF THE RANK-FREQUENCY DISTRIBUTION OF ELEMENTS IN A LARGE DATABASE

W.J. PHILLIPS AND M.A. SHEPHERD
TECHNICAL UNIVERSITY OF NOVA SCOTIA
HALIFAX, CANADA B3J 2X4

ABSTRACT

This research is concerned with analyzing the rank-frequency distribution of various elements within a 128,000 document database. Initially we were interested in knowing which elements follow Zipf's Law. In its simplest form, Zipf's Law says that the frequency of occurrence of a term in a database is (approximately) inversely proportional to its rank, where the most frequent term is rank 1.

A regression of the log of frequency versus log of rank seems to indicate that all elements considered follow Zipf's Law. A goodness of fit test, however, shows that one of these elements does not follow Zipf's Law. This show that testing for Zipf's Law cannot be based on the quality of the regression alone.

RANK-FREQUENCY DISTRIBUTION

INTRODUCTION

This paper presents the results of an investigation into the rank-frequency distribution of certain elements in a large bibliographic database. The investigation was undertaken initially more on the basis of curiosity about the database than on curiosity about the rank-frequency distribution. While satisfying our curiosity concerning the database, the investigation awakened a curiosity about the rank-frequency distribution.

The Technical University of Nova Scotia received a large portion of the OON database from the Canada Institute for Scientific and Technical Information (CISTI) for the purpose of research in the field of Information Science. The OON database contains records of conferences, monographs, and technical reports in science, technology, and medicine held by CISTI and its branch facilities. The database was received in PCAN DOBIS MARC format.

As this database supports a number of researh projects, it was felt that various characteristics of the database should be known. Therefore, certain elements of approximately 128,000 English language monograph records were tested to see if they conform to the Zipf distribution. The elements tested were: keywords drawn from the titles, Library of Congress Subject Headings, Library of Congress Classification Numbers, and author names. Not all of the elements were present in all records.

All words in the title field except those appearing on a stop list of noise words (Hunt et al., 1975) were considered keywords. The keywords were not stemmed and misspellings were not corrected.

Only topical subject headings were included. The subject headings were kept intact; i.e., phrases were not split into single terms and terms were not stemmed.

The Library of Congress Classification number and all author names, both personal and corporate, were included.

ZIPF'S LAW

For each of the four elements, authors, lcnumbers, subjects headings, and keywords, two files were prepared. The first file lists each distinct term along with its frequency x and rank r (the word of highest frequency is of rank 1). The second file lists each frequency x and the number of distinct terms of that frequency n(x).

The form of Zipf's law of interest in this paper is given by the formulas (Haitun, 1982, p. 7, formula 10):

$$x(r) = A/(r + B)^q \qquad r(x) = a/x^p - b$$

where $p = 1/q$, $b = B$ and $a = A^p$.

Let nw denote the total number of distinct terms. Let N(x) denote the number of terms of frequency less than or equal to x. It follows that $N(x) = nw - r(x+1)$. If we regard the values of x occuring in the second table as observations of a discrete random variable taking on values $x=1,2,3,\ldots$, then the cumulative distribution function F(x) is approximated by N(x)/nw. That is, if the data follows Zipf's law, the formula for F(x) is:

$$F(x) = 1 - (a/(x+1)^p - b)/nw$$

The associated mass function is $f(x) = (a/x^p - a/(x+1)^p)/nw$. This leads to the formula:

$$n(x) = a/x^p - a/(x+1)^p$$

Various authors have tested this formula for n(x). Booth (1967) tested the formula for low values of x and found reasonable agreement in four small databases. Fedorowicz (1982) tested the formula on three large databases by grouping the frequencies using powers of 2 and then performing a regression. Fedorowicz found very high correlation coefficients indicating a reasonable fit. Neither of these authors performed a goodness of fit test in the formal sense.

TESTING THE DISTRIBUTION

Since we have a formula for the cumulative distribution function it is possible to use the Kolmogorov-Smirnov goodness of fit test. The statistic for this test is $(nw)^{1/2}D$, where D is the maximum difference between the sample cumulative distribution $S(x) = N(x)/nw$ and the hypothesised cumulative distribution F(x) (Gibbons, 1971). That is :

$$D = \max \{ |S(x) - F(x)| : x=1,2,3,\ldots \}$$

Table 1 gives the critical values for $(nw)^{1/2}D$. If the value of this statistic exceeds the critical value, then the hypothesis that the data follows the Zipf Law can be rejected at that significance level.

significance	0.20	0.15	0.10	0.05	0.01
$(nw)^{1/2}D$	1.07	1.14	1.22	1.36	1.63

Table 1: Critical values for Kolmogorov-Smirnov Test

To perform the test we must have values of p, a, and b. These can be estimated by plotting ln(r+b) versus ln(x) for various values of b. The graph should approximate a straight line of slope -p and intercept ln(a). These graphs with the estimated lines appear in figures 1 to 4 in the appendix. The data bends away from the line at high values of ln(x). This is Bradford's "nucleus" (Brookes, 1984). It should be noted that this nucleus has little effect on the value of the statistic D. In any case the data cannot fall below ln(1+b). The data must fit the line very well at low values of ln(x) otherwise the statistic D will be very large. It follows that the best line is not necessarilty the regression line of ln(r+b) on ln(x) although the R2 value from this regression is a measure of how close the data is to following the Zipf distribution. Table 2 summarizes the tests for each of the four elements considered.

Element	nw	p	a	b	R^2	$(nw)^{1/2}D$
Authors	45,763	2.211	45,767	4	97%	0.41
Lcnumbers	8,458	0.997	8,608	150	89%	1.03
Subjects	15,570	0.847	15,870	300	91%	0.62
Keywords	31,940	0.765	24,530	200	97%	42.58

Table 2: Results of regression and goodness of fit test.

RESULTS

As can be seen from the values of the statistic, we cannot reject the Zipf distribution for three of the four elements. In fact, the statistic falls well beyond the 20% significance level, indicating that we cannot reject the Zipf Law for authors, lcnumbers, or for subject headings. For keywords, however, the test statistic is very large indicating that we should reject the hypothesis that the keywords follow the Zipf Law.

The statistic for the keywords is quite large because the rank predicted at $x = 1$ by the formula $r = a/x^p - b$ is quite far from nw (see figure 4). This could possibly be due to a large number of misspelled keywords which would appear as words of frequency 1. If keywords of frequency 1 are ignored in the computation of the statistic, then the value of the statistic becomes 1.08 indicating that this part of the keyword data follows the Zipf Law.

Note that the R^2 values are very high for all four elements even though the keywords do not follow the Zipf Law. A very small deviation from the line at small values of $\ln(x)$ will contribute a large amount to the statistic. This means that the R2 value from the regression cannot be used to test for the Zipf distribution. This calls into question those papers which test for the Zipf law based on regression alone.

REFERENCES

BOOTH, A. D. "A Law of Occurrences of Words of Low Frequency", Information and Control. Vol. 10, No. 4, (1967), pp. 386-393.

BROOKES, B. C. "Ranking Techniques and the Empirical Log Law", Information Processing & Management. Vol. 20, No. 1. (1984), pp. 37-46.

FEDOROWICZ, J. E. "A Zipfian Model of an Automatic Bibliographic System: An Application to MEDLINE", Journal of the American Society for Information Science. Vol. 33, No. 4. (1982) pp. 223-232.

GIBBONS, J. D. Nonparametric Statistical Inference. McGraw-Hill, 1971.

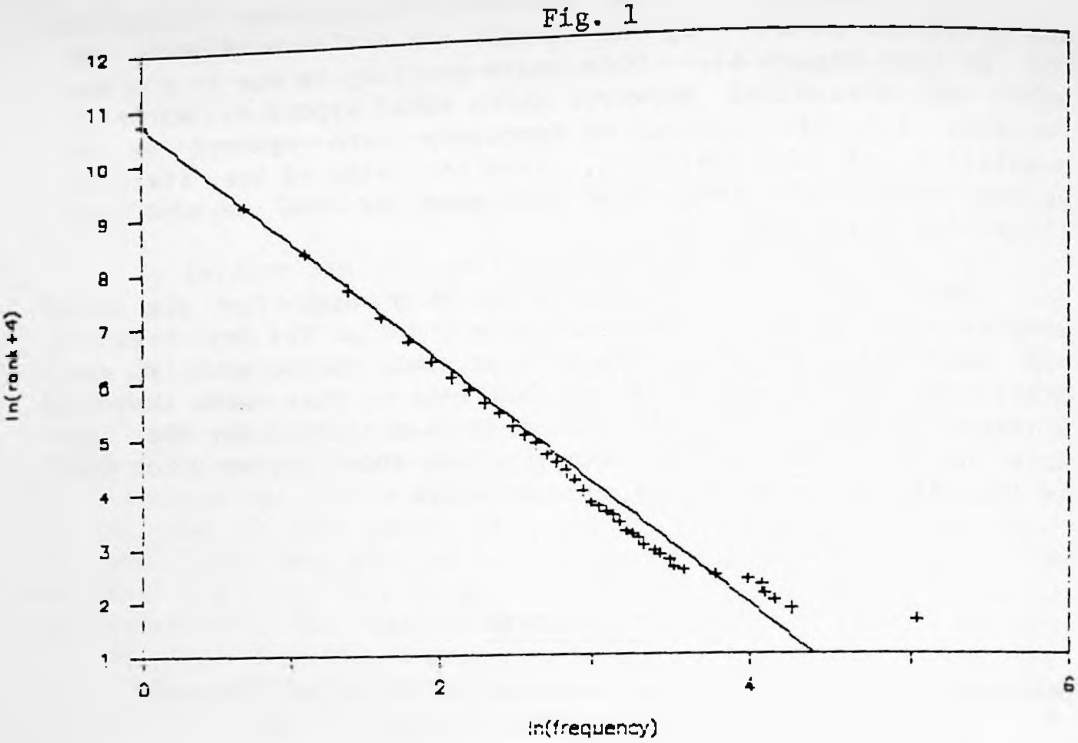
HAITUN, S. D. "Stationary Scientometric Distributions, Part I", Scientometrics. Vol. 4, No. 1. (1982), pp. 5-25.

HUNT, B.L., SNYDERMAN, M., and PAYNE, W. "Machine-Assisted Indexing of Scientific Research Summaries", in Journal of the American Society for Information Science. Vol. 26, (1975), pp. 230-236.

RANK-FREQUENCY DISTRIBUTION

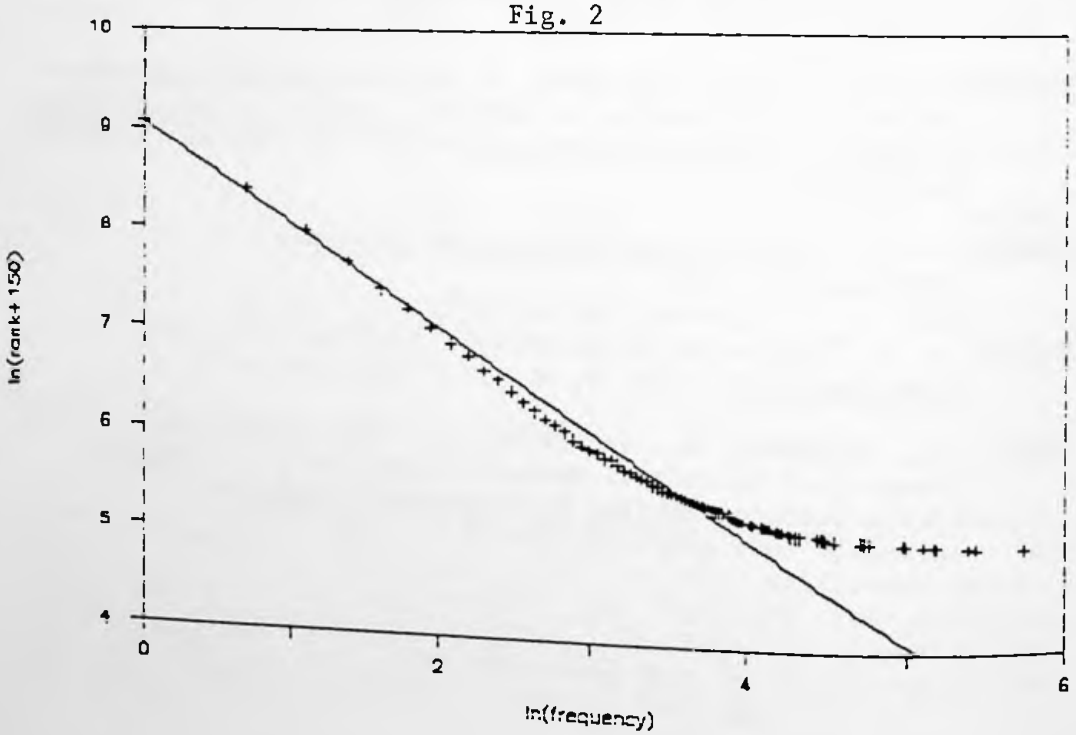
AUTHORS

Fig. 1



LIBRARY OF CONGRESS NUMBERS

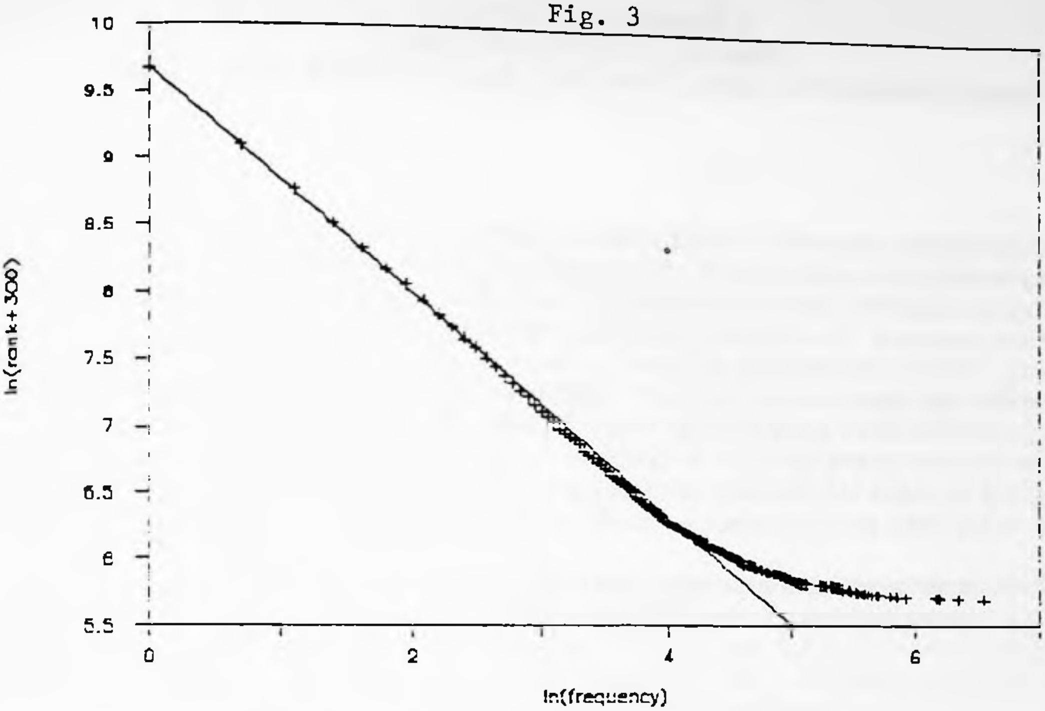
Fig. 2



RANK-FREQUENCY DISTRIBUTION

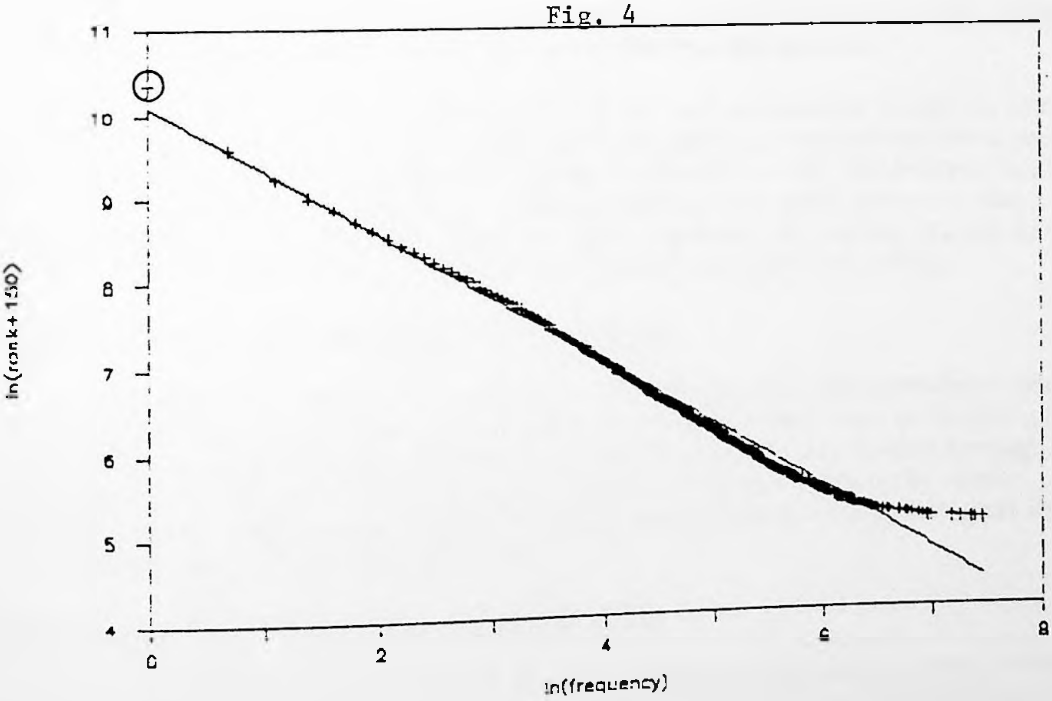
SUBJECT HEADINGS

Fig. 3



KEYWORDS

Fig. 4


ln(rank+150)
ln(frequency)