Annals of Library and Information Studies
Vol. 56, June 2009, pp.95-102

Lotka's law and authorship distribution in nutrition research in Bangladesh

S. M. Zabed Ahmed[1] and Md. Anisur Rahman[2]
[1]Professor, Department of Information Science and Library Management, University of Dhaka, Dhaka-1000, Bangladesh and Visiting Researcher (Asia Fellows Award), Institute of East Asian Studies, Thammasat University at Rangsit, Pathum Thani-12121, Thailand Email: smzahmed@yahoo.com
[2]Head, Library and Information Division, Northern University Bangladesh, Dhaka-1215, Bangladesh
Email: anis.lisbd@gmail.com

This paper examines the validity of Lotka's law to authorship distribution in the field of nutrition research in Bangladesh. A list of periodical articles on various aspects of nutrition research in Bangladesh published during 1972-2006 was compiled for analysis. Using "full productivity" of authorship, a total of 998 personal author names were identified. Lotka's law was tested using both generalized and modified forms and Kolmogorov-Smirnov goodness-of-fit tests were applied. The results suggest that author productivity distribution predicted in Lotka's generalized inverse square law is not applicable to nutrition research of Bangladesh. Using least-squares excluding high productive authors and maximum likelihood methods, Lotka's law is found to be applicable to nutrition research of Bangladesh.

Introduction

Lotka's law[1] describes the frequency of publication by authors in a given subject field. It stated that "... the number (of authors) making n contributions is about $1/n^2$ of those making one; and the proportion of all contributors, that make a single contribution, is about 60 percent." This means that out of all the authors in a given subject, about 60% publish only one article, 15% ($1/2^2$ times .60) publish two articles, 7% ($1/3^2$ times .60) publish three articles, and so on. According to Lotka, only 6% authors in a subject field produce more than ten articles. Lotka's law is often called inverse square law indicating that there is an inverse relation between the number of publications and the number of authors producing these publications.

The generalized form of Lotka's law can be expressed as $x^n y = c$, where y is the number of authors with x articles, the exponent n and constant c are parameters to be estimated from a given set of author productivity data. Lotka examined journal articles in chemistry and physics and found n values as 1.888 and 2.02 respectively; although some studies incorrectly claimed that the data only fit Lotka's law if the exponent value is exactly 2. In this paper, least-squares[2,3] and maximum likelihood[4,5] methods were applied to test the validity of Lotka's law in the field of nutrition research of Bangladesh.

Literature review

There has been considerable research conducted on the empirical validation of Lotka's law. Although many studies have confirmed the validity of the law, they often found that the exponent n is not always 2 but rather a variable value. For example, Pao[6] examined 48 datasets of author productivity covering twenty subject fields and three large research library catalogues. She found over 80% of the datasets conformed to Lotka's law in which only seven sets corroborated $n = 2$. Recent studies on various subject fields also corroborated Lotka's finding and the value of n was found to be around 2[7,8,9].

In his original studies, Lotka credited only the senior author for each contribution ignoring all co-authors. According to Potter[10], Lotka used the senior author count because multiple-authorship was less common in Lotka's time. It has been argued that ignoring all co-authors would eliminate a substantial portion of authors particularly for subjects where co-authoring is intense. A number of studies, however, showed that using total or even fractional counting of authorship lead to a breakdown of Lotka's law.[11,12] Consequently, other interpretations and formulations of Lotka's law appeared in the literature.

Pao[13] described a least-squares method for testing Lotka's law. She suggested the procedures for computing values of the exponent n and constant c and the subsequent Kolmogorov-Smirnov (K-S) goodness-of-fit

test for conformity. Some weaknesses of Pao's methodology are reported in the literature such as the fact that the least-squares approach gives acceptable results only if author data are truncated.

Nicholls[14] applied Lotka's law using all authors (without truncation) and the maximum likelihood (ML) approach to estimate parameters. This paper convincingly showed that the ML method is generally better. Newman[15] noted that the maximum likelihood is a good method and that there is a tendency for least-squares fits to overestimate the slope of the power law since the statistical fluctuations in the logarithms of the data are greater in the downward direction than in the upward one. Following Nicholls's methodology, Rousseau and Rousseau[16] developed a straightforward computer program called *Lotka* for determining the best fitting parameters for a Lotka distribution. The program also applies a Kolmogorov-Smirnov (K-S) test for conformity.

Newby *et al.*[17] used Lotka's law to test the productivity of programmers in open source software development. Programmers are considered authors and software is considered as publication. They found predicted n value as 2.82 which they claimed to have "best fit" as measured by total squared prediction error. In his observation on this paper, Burrell[18] however, pointed out the need for normalizing values when comparing two distributions so that the sum of the observed and expected values are the same, not so that the first value agrees with the first observed value.

In a recent paper, Petek[19] studied personal name headings in the Slovenian online catalogue COBIB. Pao's methodology was used by the author taking only senior author count and excluding the most prolific authors. It was found that the value of the exponent $n = 2.2656$ and the constant $c = 0{,}6890$ for COBIB. Using a K-S test, the study concluded that Lotka's law holds for the occurrences of personal name headings in COBIB. The observed distribution in COBIB was also tested against the inverse square law using the exponent n value as 2; it was found that the COBIB data do not conform to Lotka's law.

Lotka's law seems to be very resilient feature of intellectual productivity in many different subject fields. In an earlier paper,[20] Lotka's law was tested in nutrition literature of Bangladesh by simply examining the observed and theoretical values with $n = 2$. In this paper, least-squares (LS) and maximum likelihood (ML) methods were applied to test the validity of Lotka's law to nutrition research of Bangladesh. Follow-up Kolmogorov-Smirnov (K-S) tests were conducted for conformity of the results.

Objectives of the study

This paper aims to analyze authorship distribution in the field of nutrition research in Bangladesh with the following objectives:

- To examine the validity of Lotka's law, both in generalized and modified forms, using "full productivity" of authorship, and
- To undertake K-S statistics for the conformity of the results obtained by these methods.

Methodology

This study covers only periodical articles published during 1972-2006 on various aspects of nutrition in Bangladesh. The articles were primarily identified via National Library of Medicine's (NLM) PubMed using *Bangladesh* AND *nutrition* as MeSH terms. Several local journals also publish peer review articles on nutrition; some of these journals were not indexed by PubMed. To incorporate those articles, the contents page(s) of those journal issues were checked to identify papers pertaining to nutrition of Bangladesh and they were then included for analysis. The references cited by the authors in their published papers were also checked and articles which have not been included earlier were added to make this study as comprehensive as possible. The number of authors contributing one, two, or more articles each was counted manually. This study used "full productivity" of authorship, i.e., authors were given full credit for every publication in which his or her name appears.

This paper uses the least-squares methodology described by Pao.[21] The n value is calculated by this method using the following formula:

$$n = \frac{N\sum XY - \sum X \sum Y}{N\sum X^2 - (\sum X)^2} \quad \ldots (1)$$

where N = number of pairs of data; X = logarithm of articles (x); and Y = logarithm of authors (y).

The value of constant c is calculated using the following formula:

Table 1 — Frequency distribution of research contributions

No. of articles	No. of authors observed	Percentage of authors	Total no. of contributions
1	639	64.03	639
2	169	16.93	338
3	53	5.31	159
4	44	4.41	176
5	17	1.70	85
6	12	1.20	72
7	10	1.00	70
8	11	1.10	88
9	7	0.70	63
10	6	0.60	60
11	6	0.60	66
12	4	0.40	48
13	2	0.20	26
14	4	0.40	56
15	3	0.30	45
16	1	0.10	16
18	3	0.30	54
19	1	0.10	19
21	1	0.10	21
25	1	0.10	25
26	1	0.10	26
31	1	0.10	31
32	2	0.20	64
Total	998	100	2247

$$c = 1 / \sum_{1}^{P-1} \frac{1}{x^n} + 1/(n-1)(P^{n-1}) + 1/2*P^n + n/24*(P-1)^{n+1} \quad \ldots (2)$$

$\sum_{1}^{P-1} \frac{1}{x^n}$ = obtained by summing the first 19 terms of $\frac{1}{x^n}$

with x = 1, 2, 3, ... 19

here, P = 20; n = value obtained using formula (1); and x = number of articles.

This paper also applies maximum likelihood (ML) method to test Lotka's law for the nutrition research of Bangladesh. The best-known fitting ML method currently available is a computer program called *Lotka* by Rousseau & Rousseau.[22] It offers two columns for data input: source and production. Once the data are properly entered, the program returns the "best fitting" values of β (the Lotka exponent) and C for the dataset.

It should be noted here that obtaining a "best fit" does not guarantee that the fitted distribution is in fact a good fit in statistical terms. To assess that one needs to perform an accepted statistical test. Pao[23], Nicholls[24] and Burrell[25] suggested using Kolmogorov-Smirnov (K-S) test, a goodness-of-fit statistical test, to assert that the observed author productivity distribution is not significantly different from a theoretical distribution. This test is based on the maximum absolute difference between the observed and theoretical cumulative frequency distributions.

The K-S critical value at 5% level of significance is calculated as $1.36/\sqrt{\sum y}$, where $\sum y$ is the total number of authors under study. If the absolute maximum difference (D_{max}) is less than the K-S critical value, then the null hypothesis is accepted that the observed and

Table 2 — Calculation of exponent n for nutrition research

No. of articles (x)	No. of authors observed (y)	Log no. of articles (X)	Log no. of authors (Y)	XY	X^2
1	639	0.0000	6.4599	0.0000	0.0000
2	169	0.6931	5.1299	3.5558	0.4805
3	53	1.0986	3.9703	4.3618	1.2069
4	44	1.3863	3.7842	5.2460	1.9218
5	17	1.6094	2.8332	4.5599	2.5903
6	12	1.7918	2.4849	4.4524	3.2104
7	10	1.9459	2.3026	4.4806	3.7866
8	11	2.0794	2.3979	4.9863	4.3241
9	7	2.1972	1.9459	4.2756	4.8278
10	6	2.3026	1.7918	4.1257	5.3019
11	6	2.3979	1.7918	4.2965	5.7499
12	4	2.4849	1.3863	3.4448	6.1748
13	2	2.5649	0.6931	1.7779	6.5790
14	4	2.6391	1.3863	3.6585	6.9646
15	3	2.7081	1.0986	2.9751	7.3335
16	1	2.7726	0.0000	0.0000	7.6872
18	3	2.8904	1.0986	3.1754	8.3542
19	1	2.9444	0.0000	0.0000	8.6697
21	1	3.0445	0.0000	0.0000	9.2691
25	1	3.2189	0.0000	0.0000	10.3612
26	1	3.2581	0.0000	0.0000	10.6152
31	1	3.4340	0.0000	0.0000	11.7923
32	2	3.4657	0.6931	2.4023	12.0113
Total	998	52.9279	41.2484	61.7744	139.2123

theoretical distributions are the same. Kolmogorov-Smirnov test at 5% significance level was used to obtain "best fit" for the dataset.

Results of the study

Table 1 shows frequency distribution of author productivity in the field of nutrition research in Bangladesh. Of the 998 unique author names, 639 (64%) produced one article, 169 (17%) produced two articles and so forth. The number of authors who produced more than 10 articles is quite small (only 3%).

The estimated value of n for the dataset is calculated using formula (1). The n value in the field of nutrition in Bangladesh is 1.9035 for all author data. Table 2 shows the calculation of exponent n for overall author productivity data. Figure 1 shows the plotted fitted straight line through the dataset.

$$n = \frac{N\sum XY - \sum X \sum Y}{N\sum X^2 - (\sum X)^2}$$

$$= \frac{23 * 61.7744 - 52.9279 * 41.2484}{23 * 139.2123 - 52.9279^2}$$

$$= -762.3800/400.5218$$

$$= -1.9035$$

The least-squares method is used to estimate the best-fitting value for the slope of a regression line which is the exponent n for Lotka's law[26]. The slope is usually calculated excluding high productive authorship from the dataset. Since values of the slope change with different number of author data, several computations of n were made. Table 3 shows different values of n for different

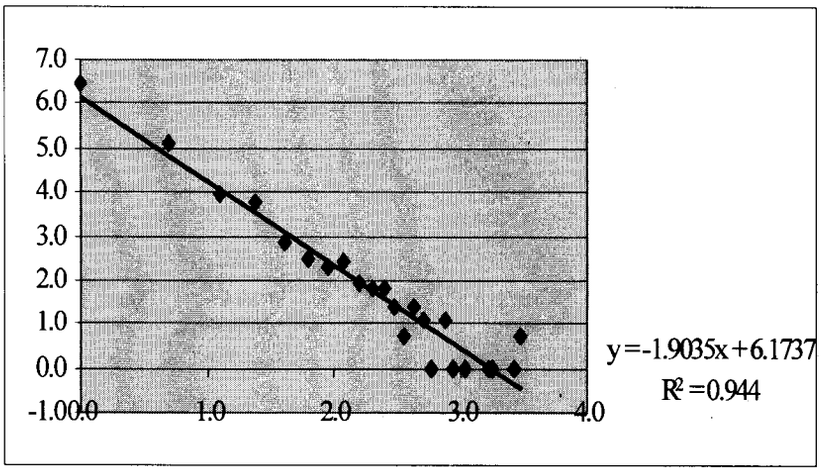

Fig. 1 — Fitted linear line for authorship distribution in nutrition research of Bangladesh

number of authors. The n value as 2.058 provides the best-fitting value for the dataset.

The n value was also calculated by maximum likelihood method using *Lotka* program. The β-value (the Lotka exponent) is 2.1758 for authorship data. For comparison, Table 4 shows the standardized fitted distribution using estimated values LS $n = 1.9035, 2.058$ and ML $n = 2.1756$.

The constant c for the dataset is calculated using formula (2). Different values of n produce different values for constant c. The c-value was first calculated using $n = 1.9035$.

$$c = 1/\sum_{1}^{P-1}\frac{1}{x^n} + 1/(n-1)(P^{n-1}) + 1/2*P^n + n/24*(P-1)^{n+1} \quad .. (2)$$

$$= 1/\sum_{1}^{19}\frac{1}{x^{1.9035}} + 1/0.9035*20^{0.9035} + 1/2*20^{1.9035} + 1.9035/24*19^{2.9035}$$

$$= 1/1.6701 + 0.0739 + 0.0017 + 0.0000$$

$$= 1/1.7457$$

$$= 0.5729$$

The calculated value of the constant c is 0.6275 for $n = 2.058$. The ML fitted distribution in *Lotka* program returned C value 0.6639 for the dataset.

Table 3 — Different values of exponent n for nutrition research of Bangladesh

No. of pairs (N)	n-value
9	2.0810
10	2.0580
11	2.0133
12	2.0115
13	2.0730
14	2.0281
15	2.0081
16	2.0812
18	2.0222
19	2.0571
21	2.0700
25	2.0522
26	2.0324
31	1.9884
32	1.9035

This study first used Kolmogorov-Smirnov (K-S) test with observed data against Lotka's inverse square law with exponent $n = 2$. If it is accepted that the proportion of all authors making a single contribution is about 60%, then the c value can be theoretically computed by the simple formula: $6/\pi^2$. Table 5 below shows the K-S test results with $n = 2$.

Table 4 — Fitted Lotka distributions with LS n = 1.9035, 2.058 and ML n = 2.1756

No. of articles	No. of authors observed	Expected with LS n = 1.9035	LS n = 1.9035 (standardized)	Expected with LS n = 2.058	LS n = 2.058 (standardized)	Expected with ML n = 2.1756	ML n = 2.1756 (standardized)
1	639	479.9586	595.6096	639	641.3486	639	673.1651
2	169	128.2901	159.2029	153.4550	154.0191	141.4425	149.0050
3	53	59.2930	73.5803	66.6170	66.8619	58.5431	61.6732
4	44	34.2912	42.5540	36.8520	36.9875	31.3083	32.9822
5	17	22.4241	27.8274	23.2820	23.3676	19.2673	20.2975
6	12	15.8487	19.6676	15.9980	16.0568	12.9585	13.6514
7	10	11.8184	14.6662	11.6490	11.6918	9.2663	9.7617
8	11	9.1658	11.3744	8.8500	8.8825	6.9301	7.3006
9	7	7.3249	9.0899	6.9450	6.9705	5.3635	5.6503
10	6	5.9938	7.4381	5.5911	5.6117	4.2648	4.4928
11	6	4.9993	6.2040	4.5953	4.6122	3.4661	3.6515
12	4	4.2363	5.2570	3.8419	3.8560	2.8684	3.0217
13	2	3.6376	4.5141	3.2584	3.2704	2.4099	2.5388
14	4	3.1590	3.9202	2.7975	2.8078	2.0511	2.1608
15	3	2.7702	3.4377	2.4272	2.4361	1.7652	1.8596
16	1	2.4500	3.0403	2.1253	2.1331	1.5340	1.6160
18	3	1.9579	2.4297	1.6678	1.6740	1.1872	1.2507
19	1	1.7664	2.1921	1.4922	1.4977	1.0555	1.1119
21	1	1.4600	1.8118	1.2144	1.2189	0.8489	0.8943
25	1	1.0477	1.3001	0.8483	0.8514	0.5810	0.6120
26	1	0.9723	1.2066	0.7825	0.7854	0.5334	0.5620
31	1	0.6957	0.8633	0.5449	0.5469	0.3638	0.3833
32	2	0.6549	0.8127	0.5104	0.5123	0.3395	0.3577
Total	998	804.2159	998	994.3453	998	947.3486	998

Table 5 — Kolmogorov-Smirnov test for n = 2

No. of articles	Observed frequency of authors	Observed cumulative frequency	Theoretical frequency of authors	Theoretical cumulative frequency	Difference
1	0.6403	0.6403	0.6079	0.6079	0.0324
2	0.1693	0.8096	0.1520	0.7599	0.0498
3	0.0531	0.8627	0.0675	0.8274	0.0353
4	0.0441	0.9068	0.0380	0.8654	0.0414
5	0.0170	0.9239	0.0243	0.8897	0.0341
6	0.0120	0.9359	0.0169	0.9066	0.0293
7	0.0100	0.9459	0.0124	0.9190	0.0269
8	0.0110	0.9569	0.0095	0.9285	0.0284
9	0.0070	0.9639	0.0075	0.9360	0.0279
10	0.0060	0.9700	0.0061	0.9421	0.0279
11	0.0060	0.9760	0.0050	0.9471	0.0288
12	0.0040	0.9800	0.0042	0.9513	0.0286
13	0.0020	0.9820	0.0036	0.9549	0.0270
14	0.0040	0.9860	0.0031	0.9580	0.0279
15	0.0030	0.9890	0.0027	0.9607	0.0282
16	0.0010	0.9900	0.0024	0.9631	0.0269
18	0.0030	0.9930	0.0019	0.9650	0.0280
19	0.0010	0.9940	0.0017	0.9667	0.0273
21	0.0010	0.9950	0.0014	0.9681	0.0269
25	0.0010	0.9960	0.0010	0.9690	0.0270
26	0.0010	0.9970	0.0009	0.9699	0.0271
31	0.0010	0.9980	0.0006	0.9706	0.0274
32	0.0020	1.0000	0.0006	0.9712	0.0289

Table 6 — Kolmogorov-Smirnov test for LS n = 1.9035

No. of articles	Observed frequency of authors	Observed cumulative frequency	Theoretical frequency of authors	Theoretical cumulative frequency	Difference
1	0.6403	0.6403	0.5729	0.5729	0.0674
2	0.1693	0.8096	0.1531	0.7260	0.0836
3	0.0531	0.8627	0.0708	0.7968	0.0659
4	0.0441	0.9068	0.0409	0.8377	0.0691
5	0.0170	0.9239	0.0268	0.8645	0.0594
6	0.0120	0.9359	0.0189	0.8834	0.0525
7	0.0100	0.9459	0.0141	0.8975	0.0484
8	0.0110	0.9569	0.0109	0.9085	0.0485
9	0.0070	0.9639	0.0087	0.9172	0.0467
10	0.0060	0.9700	0.0072	0.9244	0.0456
11	0.0060	0.9760	0.0060	0.9303	0.0456
12	0.0040	0.9800	0.0051	0.9354	0.0446
13	0.0020	0.9820	0.0043	0.9397	0.0422
14	0.0040	0.9860	0.0038	0.9435	0.0425
15	0.0030	0.9890	0.0033	0.9468	0.0422
16	0.0010	0.9900	0.0029	0.9497	0.0403
18	0.0030	0.9930	0.0023	0.9521	0.0409
19	0.0010	0.9940	0.0021	0.9542	0.0398
21	0.0010	0.9950	0.0017	0.9559	0.0391
25	0.0010	0.9960	0.0013	0.9572	0.0388
26	0.0010	0.9970	0.0012	0.9583	0.0387
31	0.0010	0.9980	0.0008	0.9592	0.0388
32	0.0020	1.0000	0.0008	0.9599	0.0401

The maximum difference (D_{max}) between the observed and theoretical values with n = 2 is 0.0498 which is greater than the critical value of 0.0435 at 5% level of significance. Therefore, the null hypothesis is rejected and concluded that the dataset does not follow Lotka's generalized inverse square law.

Again, the K-S statistic is performed to test the observed and estimated values of LS n = 1.9035 (see Table 6 below). The maximum absolute difference D_{max} is 0.836 which also falls outside the critical value of 0.0435 at 5% significance level. However, Kolmogorov-Smirnov statistics for n = 2.058 found D_{max} value 0.0314 which is within the critical value at the 5% significance level. The *Lotka* program for K-S statistics for ML distribution is 0.0236 which is also below the 5% critical value of significance and hence, both should be accepted as appropriate models for the dataset.

Conclusion

Lotka's law of author productivity is regarded as one of the classical laws of bibliometrics. This study showed that Lotka's generalized inverse square law using "full productivity" of authorship is not applicable to nutrition literature of Bangladesh. Using least-squares method, this study found n = 1.9035 and c = 0.5729 for overall data. The K-S statistics at 5% level indicate that Lotka's law is not valid for the nutrition research of Bangladesh. However, Lotka's law holds with n = 2.058 and c as 0.6275 when the high productive authors are excluded from the data. The ML fitted distributions also follows Lotka's law.

This is a preliminary study on authorship distributions in the field of nutrition research of Bangladesh; this study may trigger more such research for the purpose of evaluating nutrition research in the country. Future research should be directed towards understanding authorship distributions within various sub-fields of nutrition, authorship patterns in monographs and other publication types, collaborative authorship, author affiliation, oriental name headings, etc. Such studies would be useful in understanding the development of nutrition research in Bangladesh.

References

1. Lotka A J, The frequency distribution of scientific productivity, *Journal of the Washington Academy of Sciences*, 16(2) (1926) 317-323.
2. Pao M L, Lotka's law: a testing procedure, *Information Processing & Management*, 21(4), (1985) 305-320.
3. Pao M L, An empirical examination of Lotka's Law, *Journal of the American Society for Information Science*, 37(1) (1986) 26-33.
4. Nicholls P T, Bibliometric modelling processes and the empirical validity of Lotka's law, *Journal of the American Society for Information Science*, 40(6) (1989) 379-385.
5. Rousseau B and Rousseau R, Lotka: a program to fit a power law distribution to observed frequency data, *CYBERmetrics*, 4(1) (2000) paper 4.
6. Pao M L, *op. cit.* (1985).
7. Patra S K and Mishra S, Bibliometric study of bioinformatics literature, *Scientometrics*, 67(3) (2006) 477-489.
8. Patra S K and Chand P, HIV/AIDS research in India: a bibliometric study, *Library & Information Science Research*, 29(1) (2007) 124-134.
9. Petek M, Personal name headings in COBIB: testing Lotka's law, *Scientometrics*, 75(1) (2008) 175-188.
10. Potter W G, Lotka's law revisited. *Library Trends*, 30(1) (1981) 21-39.
11. Rousseau R, Breakdown of the robustness property of Lotka's law: The case of adjusted counts for multiauthorship attribution, *Journal of the American Society for Information Science*, 43(10) (1992) 645-647.
12. Kretschmer H and Rousseau R, Author inflation leads to a breakdown of Lotka's law, *Journal of the American Society for Information Science and Technology*, 52(8) (2001) 610-614.
13. Pao M L, *op. cit.* (1985).
14. Nicholls P T, *op. cit.*
15. Newman M E J, Comments to the article by Rousseau and Rousseau (2000), *CYBERmetrics*, 4(1), (2000) (1-2).
16. Rousseau B and Rousseau R, *op. cit.*
17. Newby G B, Greenberg J and Jones P, Open source software development and Lotka's law: bibliometric patterns in programming, *Journal of the American Society for Information Science and Technology*, 54(2) (2003) 169-178.
18. Burrell Q L, Fitting Lotka's law: some cautionary observations on a recent paper by Newby *et al.* (2003), *Journal of the American Society for Information Science and Technology*, 55(13), (2004) 1209-1211.
19. Petek M, *op. cit.*
20. Ahmed S M Z and Rahman M A, Nutrition literature of Bangladesh: a bibliometric study, *Malaysian Journal of Library and Information Science*, 13(1) (2008) 35-43.
21. Pao M L, *op. cit.* (1985).
22. Rousseau B and Rousseau R, *op. cit.*
23. Pao M L, *op. cit.* (1985).
24. Nicholls P T, *op. cit.*
25. Burrell Q L, The Kolmogorov-Smirnov test and rank-frequency distributions, *Journal of the American Society for Information Science*, 45(1) (1994) 59.
26. Pao M L, *op. cit.* (1985).