Empirical probability distribution of journal impact factor and over-the-samples stability in its estimated parameters

SK Mishra

North-Eastern Hill University, Shillong (India)

20. February 2010

Empirical Probability Distribution of Journal Impact Factor and Over-the-Samples Stability in its Estimated Parameters

SK Mishra
Dept. of Economics
North-Eastern Hill University
Shillong (India)
Contact: mishrasknehu@yahoo.com

Introduction: Do Journal Impact Factors (JIF) follow any specific probability distribution? This question has been investigated by many researchers. There is no uniformity or generality in their findings, which pertain to negative exponential (Brookes, 1970), combination of exponentials (Avramescu, 1979), Poisson (Brown, 1980), generalized inverse Gaussian-Poisson (Sichel, 1985; Burrell and Fenton, 1993), lognormal (Matricciani, 1991; Egghe and Rao, 1992), Weibull (Hurt and Budd, 1992; Rousseau and West-Vlaanderen, 1993), gamma (Sahoo and Rao, 2006), negative binomial (Bensman, 2008), approximately normal (Stringer et al., 2008), normal (Egghe, 2009), generalized Waring (Glänzel, 2009; see Panaretos and Xekalaki, 1986; Irwin, 1975), etc. It is also believed (Wikipedia, 2010) that JIFs should follow the Bradford (or Pareto) distribution, although, following the arguments of Tol (2009), JIFs are subject to the Mathew effect and, therefore, their distribution would have the tail thicker than that of the Bradford (Pareto) distribution. JIF distributions are always asymmetric and non-mesokurtic. Mishra (2010) found that in case of most of the major discipline groups (such as biology, chemistry, economics and statistics, engineering, physics, psychology and social sciences) Burr-XII, Dagum, or Johnson SU distribution are best fit to $\log_{10}$(JIF) data for 2006.

The data on JIFs provided by Thomson Scientific can only be considered as a sample since they do not cover the entire universe of those documents that cite an intellectual output (paper, article, etc) or are cited by others. Then, questions arise if the empirical distribution (best fit to the JIF data for any particular year) really represents the true or universal distribution, are its estimated parameters stable over the samples and do they have some scientific interpretation? It may be noted that if the estimated parameters do not exhibit stability over the samples (while the sample size is large enough), they cannot be scientifically meaningful, since science is necessarily related with a considerable degree of regularity and predictability. Stability of parameters is also a precondition to other statistical properties such as consistency. If the estimated parameters lack in stability and scientific meaning, then the empirical distribution, howsoever fit to data, has little significance.

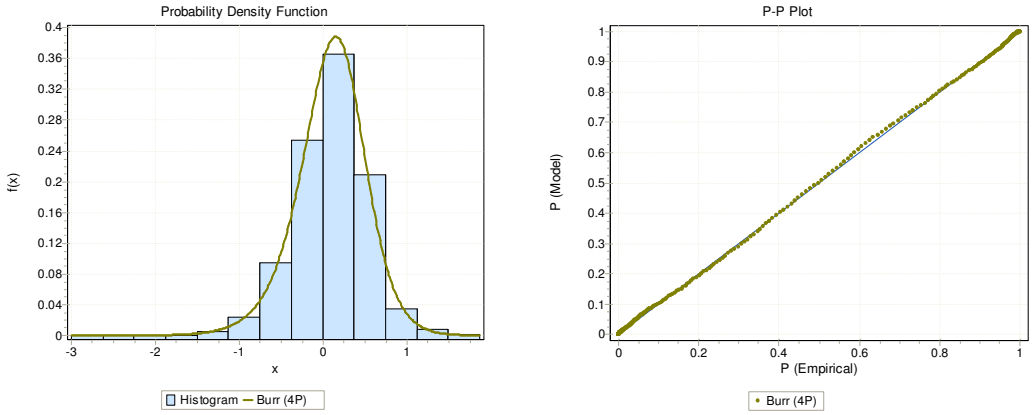
For a given year, the JIF data provided by Thomson Scientific makes a sample of a fixed size. This entire sample cannot be used to study over-the-samples stability of the parameters of empirical distribution(s). One has to draw smaller samples (better called the sub-samples) from it. That is to say that if for a given year the entire body of data on JIF is a set S of n elements, $x_1, x_2,\ldots,x_n$ pertaining to n journals, then a subsample s_1 of size $n_1<n$ is a proper subset of the set S (that is, $s_1 \subset S$). Moreover, for the purpose of random sampling, the elements of the sub-sample s_1 are randomly chosen from the elements of the set S. If n_1 is sufficiently smaller than n, then from S one can draw many sub-samples, $s_1, s_2,\ldots, s_m$. Any suitable statistical distribution can be fitted to the data in these samples to obtain its estimated parameters. Obviously, there will be sampling variations in the estimated parameters. If the sample variations are within the reasonable limits, the estimated parameters are stable over the sub-samples.

The Objectives: Our objective in this study is, first, to study the over-the-samples stability of the estimated parameters of the statistical distributions best fit to the JIF data of the year 2008 and

secondly to choose among such best fit distributions the one that exhibits the largest degree of stability in its estimated parameters. At our disposal, we have the positive JIF values for 6545 journals. This data makes the set S of n=6545 elements. From this S we randomly (uniformly distributed) draw 30 sub-samples, s_1, s_2,..., s_m: m=30, each of the size 5000. It may be noted that these sub-samples are quite large since 5000 is about 76.39 percent of 6545. We believe that such a sizable sub-sample will sufficiently represent the sample, S. Which distributions to fit to the data? We have tried with numerous distributions such as beta, Burr (4p) - also called the Singh-Maddala distribution (Singh and Maddala, 1976), Cauchy, Chi-Squared (2p), Dagum (4p), Erlang (3p), generalized normal, error function, Frechet (3p), gamma (3p), Gen. extreme value, gen. gamma (4p), Gumbel-min, Gumbel-max, hypersecant, inv. Gaussian, Johnson-SU, Laplace, Levy (2p), logistic, log-logistic (3p), normal, Pearson-5 (3p), Pearson-6 (4p), pert, Rayleigh (2p) and Weibull. It may be noted that all these distributions, except the normal, are either asymmetric or non-mesokurtic or both. We expect the best fit distributions to be both skewed and non-mesokurtic. The goodness-of-fit of the distributions is measured by three statistics pertaining to Kolmogorov-Smirnov (KS), Anderson-Darling (AD) and Chi-squared (CS) tests.

The Findings: Three distributions that emerge the best fit are: Burr (4p), Dagum (4p) and Johnson SU. In the majority of cases either Burr (4p) or Dagum (4p) does better than Johnson SU on the criterion of KS test. However, on AD and CS tests, Johnson SU is emerges stronger than on KS test. It may be noted that AD weights the fit to the tails more and CS weights the overall fit more.

Fig.1.1: Histogram, pdf and P-P plot of Burr(4p) Distribution fitted to Log_{10}(JIF) Data (2008)



-

Fig.1.2: Histogram, pdf and P-P plot dagum(4p) Distribution fitted to Log_{10}(JIF) Data (2008)

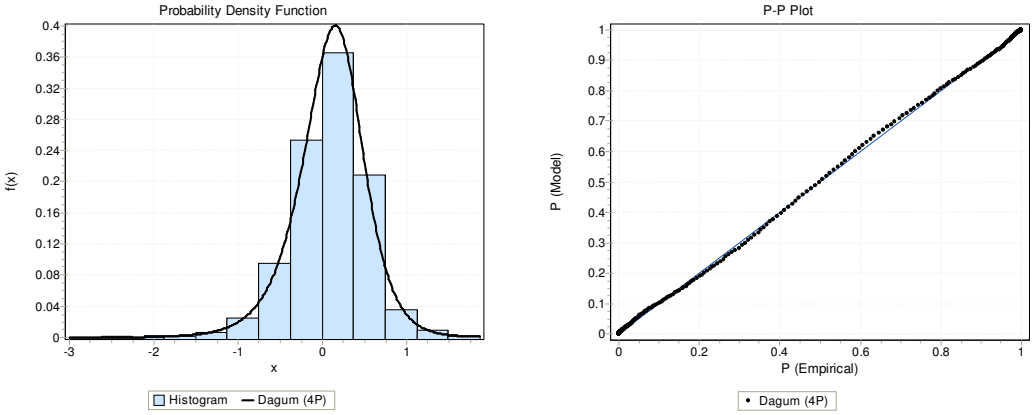


Fig.1.3: Histogram, pdf and P-P plot Johnson SU Distribution fitted to Log_{10}(JIF) Data (2008)

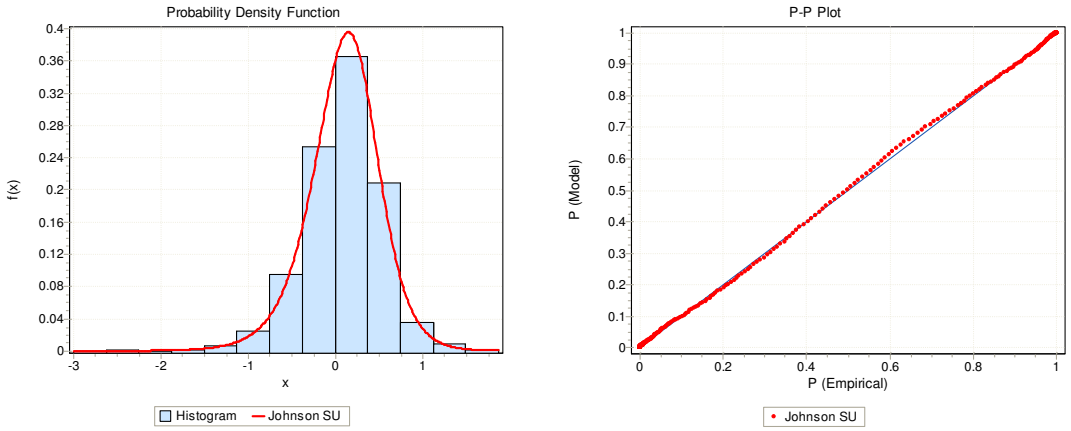


Table 1.1: Estimated Parameters of Burr and Dagum Distributions to the Sub-sample data of Log_{10}(JIF) 2008								
	Estimated Parameters of Burr (4p) Distribution				Estimated Parameters of Dagum (4p) Distribution			
s_j	k	α	β	γ	k	α	β	γ
1	1.5601	209.2700	56.9940	-56.7170	0.6851	188.3000	38.7140	-38.4780
2	1.4934	60704.0000	16429.0000	-16429.0000	0.6744	240.2000	49.3810	-49.1340
3	1.6015	180.3700	49.6700	-49.3760	0.5478	34.5510	6.5984	-6.2907
4	1.8477	29.6170	8.7244	-8.3724	0.6283	73.0410	14.5320	-14.2640
5	1.5342	755.9800	204.8900	-204.6100	0.6849	663.2000	136.1100	-135.8700
6	1.5567	4469.9000	1230.8000	-1230.5000	0.6168	60.2530	12.0130	-11.7430
7	1.5564	1957.0000	532.6000	-532.3200	0.6724	312.4100	63.5600	-63.3140
8	1.5321	322.4100	87.2190	-86.9460	0.8545	103000.0000	23613.0000	-23612.0000
9	1.5602	339.8700	92.8250	-92.5410	0.5956	50.3360	9.8366	-9.5512
10	1.5889	126.1700	34.7240	-34.4350	5.0554	405000000.0000	199000000.0000	-199000000.0000
11	1.8749	54.0710	15.9180	-15.5620	0.6834	1319.1000	272.5800	-272.3500
12	1.6302	82.8780	22.7580	-22.4600	0.5286	32.0710	5.8967	-5.5795
13	1.6588	313.1500	87.2750	-86.9700	0.6620	600.6200	122.0100	-121.7600
14	1.2365	11291.0000	2826.9000	-2826.7000	0.6462	62.6790	12.6830	-12.4240
15	1.6171	41675.0000	11345.0000	-11344.0000	0.6028	69.9830	13.5740	-13.2980
16	1.6503	175.6700	48.5690	-48.2630	0.5452	34.5250	6.5576	-6.2494
17	1.5910	275.8400	76.0920	-75.8000	0.5252	17.8070	3.4667	-3.1407
18	1.6351	3870000.0000	1060000.0000	-1060000.0000	0.6165	105.8200	20.7170	-20.4470
19	1.7040	107.1600	30.7300	-30.4060	0.6500	193.5000	39.2630	-39.0010
20	1.5872	890.5400	244.5200	-244.2300	0.6851	1341.2000	276.3300	-276.0900
21	1.5595	291.3500	79.0310	-78.7450	0.5714	36.8470	7.1053	-6.8077
22	1.7537	45.9910	13.2180	-12.8940	0.7992	391000000.0000	87700000.0000	-87700000.0000
23	1.5037	362.1900	97.5640	-97.2950	0.6142	42.5570	8.4657	-8.1928
24	1.8309	33.1470	9.7576	-9.4088	0.7180	584.5900	122.1100	-121.8800
25	1.4637	21216.0000	5712.3000	-5712.0000	0.7170	9316.6000	1955.5000	-1955.2000
26	1.4263	95223.0000	25427.0000	-25426.0000	0.6708	80.2890	16.4950	-16.2420
27	2.2972	1410000.0000	448000.0000	-448000.0000	0.5666	42.5630	8.2167	-7.9171
28	1.5292	32456.0000	8820.1000	-8819.9000	0.6480	122.9400	24.8460	-24.5890
29	1.6472	124.7900	34.6950	-34.3920	0.6629	213.4600	43.5050	-43.2510
30	1.4380	1009.8000	269.8000	-269.5500	0.7401	6620000.0000	1410000.0000	-1410000.0000
s_j = sub-sample j; j=1,2,…,30 of size 5000 randomly drawn from JIF 2008 data set, S, of size 6545.								

The illustrative fits of Burr (4p), Dagum (4p) and Johnson SU distributions to sub-sample data are presented in Fig.-1 through Fig.-3. The estimated parameters of Burr (4p) and Dagum (4p) distributions are presented in Table-1.1. Variations in the estimated parameters over the samples are conspicuous. Large standard deviations with respect to mean and confidence values at -95 and +95 percent levels presented in Table 2.1 indicate the instability of parameters over the samples. Therefore, nothing can be concluded or predicted as to the behavior of those parameters for any other sub-sample or even the sample or the universe.

The estimated parameters of Johnson SU distribution are presented in Table-1.2. Measures of central tendency and dispersion of the estimated parameters are presented in Table 2.2. The two measures of central tendency (median and mean) for all the parameters indicate that their distributions are almost symmetrical. Their standard deviations are much smaller with respect to their means. It can be easily seen that the estimated parameters of the Johnson SU distribution exhibit over-the-samples stability.

Table 1.2: Estimated Parameters of Johnson SU Distribution to the Sub-sample data of Log_{10}(JIF) 2008

s_j	γ	δ	λ	ξ	s_j	γ	δ	λ	ξ	s_j	γ	δ	λ	ξ
1	0.4388	2.0226	0.7497	0.2819	11	0.4742	1.9210	0.7094	0.2959	21	0.4309	1.9153	0.7015	0.2878
2	0.4555	1.9248	0.7122	0.2937	12	0.4329	1.9344	0.7058	0.2861	22	0.4822	1.8967	0.6922	0.3010
3	0.5042	2.1548	0.8121	0.3164	13	0.5243	1.9924	0.7360	0.3215	23	0.4214	2.0273	0.7526	0.2821
4	0.4830	1.9832	0.7346	0.3073	14	0.4246	1.9879	0.7392	0.2845	24	0.4157	1.9758	0.7301	0.2777
5	0.4837	1.9717	0.7256	0.3085	15	0.5088	1.9694	0.7165	0.3105	25	0.4133	1.9421	0.7207	0.2756
6	0.4960	1.9791	0.7362	0.3097	16	0.5018	2.0009	0.7361	0.3140	26	0.3877	1.9525	0.7273	0.2722
7	0.4854	1.9685	0.7242	0.3050	17	0.4594	1.9489	0.7177	0.2968	27	0.4664	1.8911	0.6941	0.3031
8	0.4350	1.9454	0.7121	0.2854	18	0.5268	1.9545	0.7095	0.3202	28	0.4596	1.9496	0.7213	0.2948
9	0.4568	1.9762	0.7305	0.2973	19	0.4666	1.8943	0.7001	0.3044	29	0.4823	1.9664	0.7297	0.3079
10	0.4485	1.9235	0.7126	0.2948	20	0.4622	1.9326	0.7108	0.2983	30	0.3971	1.9677	0.7300	0.2696

s_j = sub-sample j; j=1,2,…,30 of size 5000 randomly drawn from JIF 2008 data set, S, of size 6545.

.

Table 2.1: Measures of Central Tendency and Dispersion of Parameters of Burr and Dagum Distributions over Sub-samples of JIF 2008

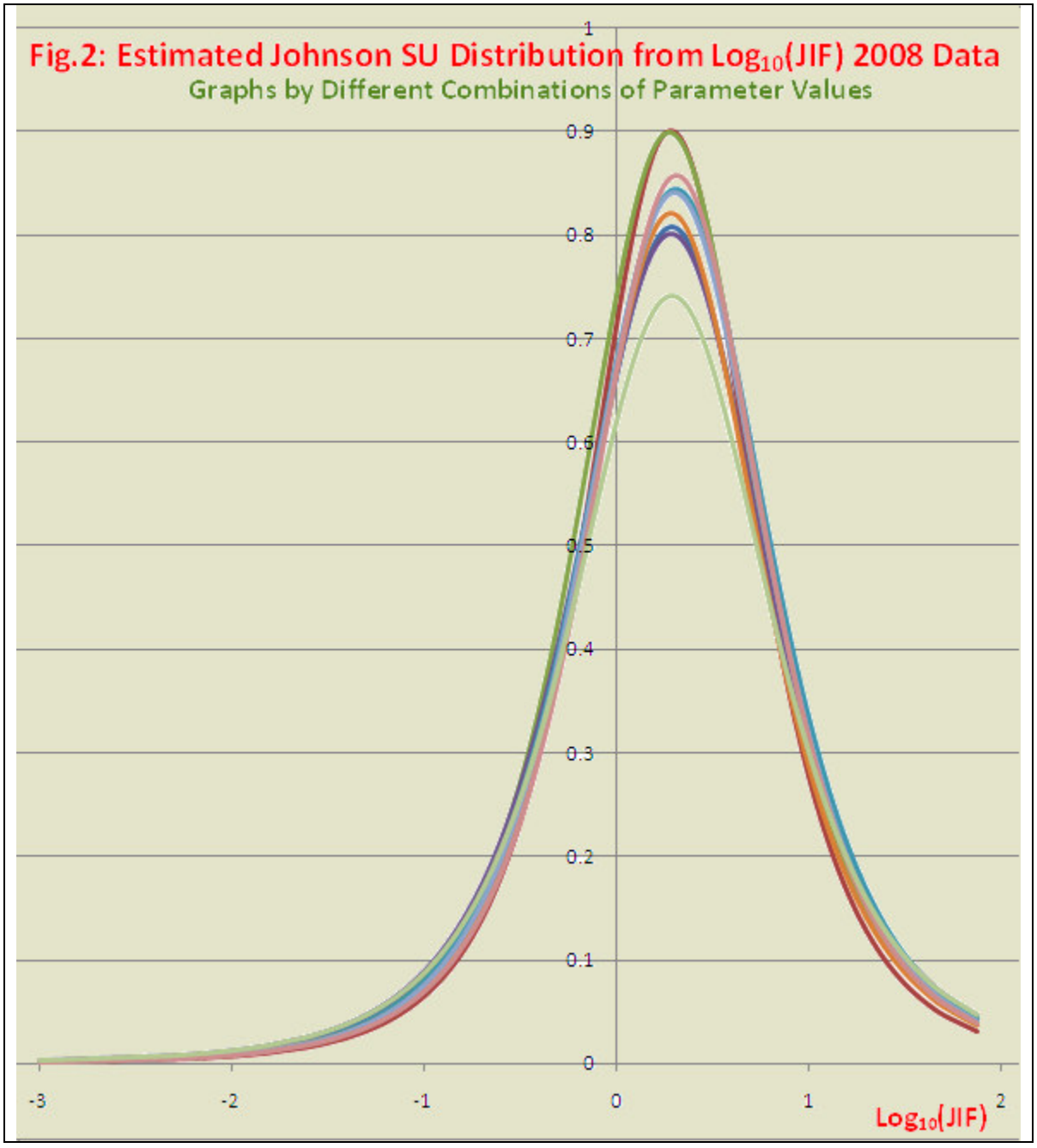
Parameter	Burr (4p) Distribution					Dagum (4p) Distribution				
	Median	Mean	Std. Dev.	Confidence -0.95%	Confidence +0.95%	Median	Mean	Std. Dev.	Confidence -0.95%	Confidence +0.95%
k	1.588	1.616	0.182	1.547	1.684	0.656	0.796	0.808	0.49	1.10
α	331.140	185157.400	741681.400	-91791.000	462105.800	155.620	26757961.310	100938728.500	-10933179.40	64449102.00
β	90.050	52729.600	206953.000	-24548.000	130007.200	31.780	9604563.436	39184290.260	-5027091.03	24236217.90
γ	-89.756	-52729.300	206953.100	-130007.000	24548.200	-31.534	-9604563.170	39184290.330	-24236217.70	5027091.32

.

Table 2.2: Measures of Central Tendency and Dispersion of Parameters of Johnson SU Distribution over Sub-samples of JIF 2008

Parameter	Median	Mean	Std. Dev.	Confidence -0.95%	Confidence +0.95%
γ	0.460860	0.460835	0.036127	0.447345	0.474325
δ	1.960450	1.962353	0.050159	1.943624	1.981083
λ	0.722745	0.724350	0.022454	0.715965	0.732735
ξ	0.297060	0.296796	0.014188	0.291498	0.302094

A graphical presentation of the estimated probability density functions (pdf) of Johnson SU distribution using various combinations of estimated values of the four parameters (Fig.2) is given below. It is seen that the distribution is nice behaved.



Conclusion: This exercise suggests that to accept the fitness of a statistical distribution to given data (in this example, the $\log_{10}$(JIF)-2008 data), it is not appropriate to depend on the goodness of fit criteria alone. Stability of parameters criterion also is a very important consideration, which may not always be satisfied by the empirically best fit statistical distribution. Secondly, the Johnson SU distribution fits best to the $\log_{10}$(JIF) data and its parameters are stable over the sub-samples. Then, will Johnson SU distribution exhibit this stability for $\log_{10}$(JIF) data in other years too? We hope it will.

References:

1. Avramescu, A. (1979) "Actuality and Obsolescence of Scientific Literature", *Journal of the American Society for Information Science*, 30(5): 296-303.
2. Bensman, S. J. (2008) "Distributional Differences of the Impact Factor in the Sciences Versus the Social Sciences: An Analysis of the Probabilistic Structure of the 2005 Journal Citation Reports", Journal of the American Society for Information Science and Technology, 59(9): 1366–1382.
3. Brookes, B. C. (1970) "The Growth, Utility, and Obsolescence of Scientific Periodical Literature", *Journal of Documentation*, 26(4): 283-294.
4. Brown, P. (1980) "The Half-life of Chemical Literature", *Journal of the American Society for Information Science*, 31(1): 61-63.
5. Burrell, Q. and Fenton, M. R. (1993) "Yes, the GIGP Really Does Work – and is Workable", *Journal of the American Society for Information Science*, 44(2): 61-69.
6. Egghe, L. (2009) "Mathematical Derivation of the Impact Factor Distribution", *Journal of Informetrics*, 3(4): 290-295.
7. Egghe, L. and Rao, I. K. (1992) "Citation Age of Data and the Obsolescence Function: Fits and Explanations", *Information Processing and Management*, 28(2): 201-217.
8. Glänzel, W. (2009) "The Multi-dimensionality of Journal Impact", *Scientometrics*, 78(2): 355-374.
9. Hurt, C. D. and Budd, J. M. (1992) "Modeling the Literature of Superstring Theory: A Case of Fast Literature", *Scientometrics*, 24(3): 471-480.
10. Irwin, J. O. (1975) The Generalized Waring Distribution. Part I, *Journal of the Royal Statistical Society. Series A (General)*, 138: 18–31.
11. Matricciani, E. (1991) "The Probability Distribution of the Age of References in Engineering Papers", *IEEE Transactions on Professional Communication*, 34(1): 7-12.
12. Mishra, S. K. (2010) "A Note on Empirical Sample Distribution of Journal Impact Factors in Major Discipline Groups", available at SSRN: http://ssrn.com/abstract=1552723
13. Panaretos, J. and Xekalaki, E. (1986) "The Stuttering Generalized Waring Distribution", *Statistics & Probability Letters*, 4(1986) 313-318.
14. Rousseau, R. and West-Vlaanderen, K. I. H. (1993) "A Note on Maximum Impact Factors", Available at http://www.cais-acsi.ca/proceedings/1993/Rousseau_1993.pdf
15. Sichel, H.S. (1985) "A Bibliometric Distribution which Really Works", *Journal of the American Society for Information Science*, 36: 314-321.
16. Sahoo, B. B. and Rao, I. K. R. (2006) "A Distribution of Impact Factors of Journals in the Area of Software: An Empirical Study", Information Processing & Management, 42(6): 1465-1470.
17. Singh, S. K., and Maddala, G. S. (1976) "A Function for Size Distribution of Incomes", *Econometrica*, 44(5): 963-970.
18. Stringer, M. J., Sales-Pardo, M. and Amaral, L. A. N. (2008) "Effectiveness of Journal Ranking Schemes as a Tool for Locating Information", *PLoS ONE* 3(2): e1683. doi:10.1371/ journal.pone.0001683: 1-8.
19. Tol, R.S.J. (2009). "The Matthew effect defined and tested for the 100 most prolific economists", *Journal of the American Society for Information Science and Technology*, 60(2): 420-426.
20. Wikipedia (2010) "Impact Factor", available at http://en.wikipedia.org/wiki/Impact_factor

Appendix
Algebraic form of the pdf of Burr (4p), Dagum (4p) and Johnson SU Distributions

i. Burr-XII Distribution: It is also known as 4-parameter generalized Beta-II distribution with unit shape parameter, Singh-Maddala distribution (Singh and Maddala, 1976) as well as the Pareto-IV distribution (Kleiber and Kotz, 2003). With the support random variable $x:\gamma \le x < +\infty$, the probability density function (pdf) of Burr 4-parameters (4p) distribution is given as:

$$f(x) = \frac{\alpha k \left(\frac{x-\gamma}{\beta}\right)^{\alpha-1}}{\beta\left(1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right)^{k+1}}, \text{ where}$$

$k, \alpha > 0$ are the two shape parameters
$\beta > 0$ is the scale parmeter
γ is the location parameter
If γ=0, then the distribution is 3p

ii. Dagum (Inverse Burr-III) Distribution: With the support random variable $x:\gamma \le x < +\infty$, the probability density function (pdf) of Dagum 4-parameters (4p) distribution is given as:

$$f(x) = \frac{\alpha k \left(\frac{x-\gamma}{\beta}\right)^{\alpha k-1}}{\beta\left(1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right)^{k+1}}, \text{ where}$$

$k, \alpha > 0$ are the two shape parameters
$\beta > 0$ is the scale parmeter
γ is the location parameter
If γ=0, then the distribution is 3p

iii. Johnson SU Distribution: With the support random variable $x:-\infty < x < +\infty$, the probability density function (pdf) of Johnson SU distribution is given as:

$$f(x) = \frac{\delta}{\lambda\sqrt{2\pi}\sqrt{z^2+1}} \exp\left(-\frac{1}{2}\left(\gamma + \delta \ln\left(z+\sqrt{z^2+1}\right)\right)^2\right), \text{where}$$

ζ is the location parameter
$z = (x-\zeta)/\lambda$

γ is the shape parameter
δ>0 is another shape parameter
$\lambda > 0$ is the scale parmeter