Theory of first-citation distributions and applications

EGGHE, Leo & Ravichandra Rao, Inna Kedage (2001) Theory of first-citation distributions and applications. In: Mathematical and Computer Modelling, 34(1-2). p. 81-90.

DOI: 10.1016/S0895-7177(01)00050-4
Handle: http://hdl.handle.net/1942/784

THEORY OF FIRST-CITATION DISTRIBUTIONS AND APPLICATIONS

by

L. Egghe, LUC, Universitaire Campus, B-3590 Diepenbeek, Belgium[1]
DRTC, ISI, 8th Mile, Mysore Road, R.V. College P.O., Bangalore - 560059, India[2]
UIA, Universiteitsplein 1, B-2610 Wilrijk, Belgium
e-mail : leo.egghe@luc.ac.be

and

I.K. Ravichandra Rao, DRTC, ISI, 8th Mile Mysore Road, R.V. College P.O.,
Bangalore - 560059, India
e-mail : ikr@isibang.ac.in

ABSTRACT

The general relation between the *first-citation distribution* and the general citation-age-distribution is shown. It is shown that, if Lotka's exponent $\alpha=2$, both distributions are the same. In the light of the above results, and as a simple case, the exponential distribution and the lognormal distribution have been tested and accepted. Also the n^{th} ($n \in \mathbb{N}$) citation distribution is studied and shown to be the same as the first-citation distribution, for every $n \in \mathbb{N}$.

[1] Permanent address

[2] Research on this article has been executed while the first named author was a visiting professor in the Documentation Research and Training Centre, Indian Statistical Institute, Bangalore, India.

Key words and phrases : first-citation distribution, Lotka, exponential distribution, lognormal distribution.

I. Introduction.

The time at which an article receives its first citation is a very important moment. It changes the status of this article from unused to used. It is clear that the time t_1 of this event is a very important parameter : if t_1 is small then the article is at the front of research and/or belongs to a subject where communication between scientists is heavy (e.g. through the invisible colleges - a growing phenomenon especially since the availability of the Internet).

In our feeling, t_1 is a valuable alternative for both the classical impact factor (IF) and the immediacy index (II) as produced by the Institute for Scientific Information : the time of the first citation measures visibility as well as the time to become visible. It is an important research tool in science policy studies, yet it must be admitted that first-citation data are not readily at hand, for the time being. Of course they could be produced from the citation indexes, if only one could convince people of its importance. For the time being we will re-use the few first-citation data that are availble : data of Gupta and Rousseau (1999) in theoretical population genetics, the Motylev data appearing in Motylev (1981) on Russian language library science papers, and the JACS (Journal of the American Chemical Society) data of Rousseau (1994). To this set we will add a new data set on first-citation data in JASIS articles of 1980, followed in JASIS until now (end 1999).

The first section presents a general theory on first-citation distributions. In this theory the general citation-age distribution is considered in connection with Lotka's law on the number of papers with a certain number (say A) of citations. Here A means the total number of citations a paper receives (i.e. a diachronous study). Let us denote Lotka's law by

$$\varphi(A) = \frac{C}{A^{\alpha}} \qquad (1)$$

, where C is a constant, making sure φ is a distribution. Here $\alpha>1$ and the most classical value of α is $\alpha=2$, being the original law of Lotka - see Lotka (1926) or Egghe and Rousseau (1990) for more information on this. Let c(t) denote the general citation-age distribution, being the fraction of citations at time t (after publication of a paper). Let C(t) denote the cumulative distribution derived from t. Hence

$$C(t) = \sum_{s=0}^{t} c(s) \tag{2}$$

for discrete distributions and

$$C(t) = \int_{0}^{t} c(s)ds \tag{3}$$

for continuous ones.

Classical examples are the exponential distribution :

$$c(t) = c\, a^{t}, \tag{4}$$

where 0<a<1 *and c is a constant making* sure that c is a distribution (t can be discrete, in which case c = 1-a, or continuous, in which case c = - lna), or

$$c(t) = \frac{1}{\sqrt{2\pi}\ \sigma t}\ e^{-\left(\frac{\ln t - \mu}{\sigma}\right)^{2}} \tag{5}$$

the lognormal distribution (t continuous).

The exponential one is the most appealing and is basic to all citation studies (see e.g. Egghe and Rousseau (1990)). The parameter a denotes the aging rate, i.e. the decline in use when time goes further on, expressed by

$$a = \frac{c(t+1)}{c(t)}\ . \tag{6}$$

The lognormal distribution is - however - the more realistic one, taking into account the initial increase (for t small) of the number of citations due to the fact that a paper has an "introductory" period in which it becomes gradually more and more visible. After reaching its maximum, the decline starts, just as in the case of the exponential distribution. The values μ and σ now replace a in (4) and are the mean, respectively the standard deviation of logaritmic time, lnt. The lognormal distribution has been generally accepted and explained, see e.g. Matricciani (1991) or Egghe and Ravichandra Rao (1992).

In our first section, however, it does not matter what the exact form of c is : our results are true for general citation-age distributions, hence solving completely the relation between the first-citation distribution and the distribution c. Our main result is that, denoting by $\Phi(t_1)$ the first-citation cumulative distribution (t_1 = the time of the first citation), that

$$\Phi(t_1) = C(t_1)^{\alpha-1} \tag{7}$$

Hence, as an important general corollary we obtain that, if α=2 (the most classical value of Lotka's exponent) then

$$\Phi(t_1) = C(t_1), \tag{8}$$

i.e. the first-citation distribution is the same as the general citation-age distribution. As far as we are aware this remarkably simple fact has never been noted before (and the same goes, of course, for formula (7)).

In the light of (7) and (8), the second section is devoted to the fitting of the exponential distribution and the lognormal one to first-citation data. From the obtained results it is clear that, already in the simple case (α=2), the fits are good and hence can be accepted in the sense that concave data are fitted well by the exponential distribution and the S-shaped ones are fitted well by the lognormal distribution. In this way the Gupta and Rousseau (1999) data, the Motylev (1981) data, the Rousseau data (1994) and our JASIS data are fitted very well.

Of course, as noted in Egghe (1999), by taking $\alpha \neq 2$ (in fact $\alpha > 2$) it is possible to obtain very good fits for S-shaped data by using the exponential distribution. This case, as a special case of (7), was used in Egghe (1999) to fit Rousseau's JACS data as well as the Gupta and Rousseau data (although the S-shape is not very well apparent in this case). In all these cases (8) is an alternative, when using the cumulative lognormal for C.

The ultimate generality would be to use (7) with general α and using the lognormal distribution. In view of the above good results we doubt if this generality is necessary. It is certainly more complicated, now involving 3 parameters (μ,σ,α) and fitting powers of cumulative lognormal distributions is not belonging to standard statistical packages. All our fittings involve 2 parameters only.

II. General relation between first-citation distribution and the general citation age distribution.

Let us fix a bibliography, being a general set of documents. Each of these documents eventually receive citations. Let c(t) denote the distribution of the citations that are given to documents of this bibliography, t time units (e.g. years or months) after they are published. Let C(t) denote the cumulative distribution of c(t), e.g. (2) or (3) - we do not specify here whether t is a discrete or continuous variable. We assume C (hence c) to be the same for all documents in the bibliography.

Let $\Phi(t_1)$ denote the cumulative distribution of the documents that receive their first citation. This distribution is assumed to be conditionally w.r.t. the ever cited documents. The fraction γ of the non-cited ones will be dealt with at the end of the paper. Hence here we have that

$$\lim_{t_1 \to +\infty} \Phi(t_1) = 1 \tag{9}$$

As said in the previous section, we will assume Lotka's law for the distribution of the total number of citations per document :

$$\varphi(A) = \frac{C}{A^{\alpha}}, \tag{10}$$

, where $\alpha>1$ and where $\varphi(A)$ is the fraction of documents with A citations. Here $C=\alpha-1$ in order to make φ a distribution (in the continuous setting for the variable A, which we will adopt).

We have the following general result

Theorem 1 : The following relation between the first-citation distribution and the general citation age distribution is valid :

$$\Phi(t_1) = C(t_1)^{\alpha-1} \tag{11}$$

Proof : For each document in the bibliography that has A citations in total, we have that t_1, the time of the first citation is given by

$$AC(t_1) = 1,$$

hence

$$A = C(t_1)^{-1} \tag{12}$$

For all values A'>A we evidently have

$$A'C(t_1) > 1,$$

hence these documents belong to the ones that received their first citation before t_1. Their cumulative fraction is

$$\begin{aligned} &\int_A^\infty \varphi(A')dA' \\ &= \int_A^\infty \frac{\alpha-1}{A'^{\alpha}}dA' \\ &= A^{1-\alpha}, \end{aligned} \tag{13}$$

since α>1. Hence this also equals $\Phi(t_1)$ with A replaced by (12). Consequently

$$\Phi(t_1) = C(t_1)^{\alpha-1}. \qquad \square$$

Note that it follows from (12) that A is large iff t_1 is small. Hence the smaller t_1 the more visible the publication is, since A measures total visibility.

Corollary 2 : If $\alpha=2$, the most "classical" Lotka exponent, then

$$\Phi(t_1) = C(t_1),$$

in other words, the first-citation distribution equals the general citation age distribution.

To the best of our knowledge, this remarkably simple result has never been noted before. We can already say that - roughly speaking - if $\alpha \approx 2$ (which is so in most cases) that $\Phi \approx C$.

The above result can be extended to the (less important) case of the n^th citation distribution, i.e. the time distribution that the documents in the bibliography receive their n^th citation ($n \in N$ fixed). Of course, as was the case for n=1, not all documents will be cited n times. We therefore denote by Φ_n the conditional cumulative n^th citation distribution, i.e. w.r.t. the collection of documents that receive at least n citations. We have the following result.

Proposition 3 : For all $n \in N$, the n^th citation distribution equals the first-citation distribution :

$$\Phi_n = \Phi. \tag{14}$$

Hence

$$\Phi_n(t) = C(t)^{\alpha-1} \tag{15}$$

for all $n \in N$. The fraction (amongst all documents that are ever cited) of documents with at least n citations at time t is given by

$$\left(\frac{C(t)}{n}\right)^{\alpha-1}. \tag{16}$$

Proof : As in the proof of theorem 1 we now have

$$AC(t) = n \tag{17}$$

for the time t that a document, with A citations in total, receives n citations. Following the rest of the proof of theorem 1 exactly, gives that the fraction (amongst all documents that are ever cited) of documents with at least n citations at time t is given by $A^{\alpha-1}$, with A as in (17), hence

$$\left(\frac{C(t)}{n}\right)^{\alpha-1} . \tag{18}$$

Since

$$\lim_{t\to\infty}\left(\frac{C(t)}{n}\right)^{\alpha-1} = \frac{1}{n^{\alpha-1}} \tag{19}$$

we hence have that the conditional cumulative distribution is given by

$$\Phi(t) = C(t)^{\alpha-1} = \Phi(t). \qquad \square$$

So again, if α=2, we see that Φ_n=C for all n∈ℕ.

We return now to the case of first-citation distributions. Two important cases are the ones in which the general citation age distribution is exponential (see (4)) or lognormal (see(5)).

Let us first consider the case of the exponential distribution (4)

$$c(t) = ca^t \tag{20}$$

for continuous t (we leave the discrete case to the reader). Hence, here c = - lna. The cumulative distribution then takes the form

$$C(t) = \int_0^t -(\ln a)a^{t'}dt'$$

$$C(t) = 1\text{-}a^t \tag{21}$$

In this case theorem 1 says that

$$\Phi(t_1) = (1-a^{t_1})^{\alpha-1} \tag{22}$$

, hence we refind the result in Egghe (1999). Note that in Egghe (1999) one has multiplied by the fraction γ of uncited papers. As remarked before, we only look at cumulative distributions, i.e. conditionally w.r.t. to the collection of ever cited papers.

In Egghe (1999), distribution (22) has proved to fit very well first-citation data such as the ones of Gupta and Rousseau (1999), Motylev (1981) and Rousseau (1994). Distribution (22) is capable of fitting concave as well as S-shaped data. Indeed, as proved in Egghe (1999), (22) is concave iff $1<\alpha\leq 2$ and is S-shaped iff $\alpha>2$.

In case of the lognormal distribution, however, we have S-shapes, even for $\alpha=2$. Indeed the cumulative lognormal distribution itself is S-shaped. Let us go into this into more detail. If

$$c(t) = \frac{1}{\sqrt{2\pi}\ \sigma t}\, e^{-\left(\frac{\ln t-\mu}{\sigma}\right)^2} \tag{23}$$

its cumulative distribution is

$$C(t) = \int_0^t \frac{1}{\sqrt{2\pi}\ \sigma t}\, e^{-\frac{1}{2}\left(\frac{\ln t-\mu}{\sigma}\right)^2} dt$$

$$= \int_{-\infty}^{\frac{\ln t-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}s^2} ds$$

$$C(t) = F\left(\frac{\ln t-\mu}{\sigma}\right), \tag{24}$$

where F denotes the cumulative normal distribution. In this case theorem 1 says that

$$\Phi(t_1) = F^{\alpha-1}\left(\frac{\ln t_1 - \mu}{\sigma}\right) . \tag{25}$$

Now even for $\alpha=2$ there is an S-shape since Φ equals the cumulative lognormal distribution.

In the next section we will investigate if these simple functions (for $\alpha=2$) are capable of fitting practical first-citation data in the following sense :

$$\Phi(t_1) = 1-a^{t_1} \tag{26}$$

for concave data and

$$\Phi(t_1) = F\left(\frac{\ln t_1 - \mu}{\sigma}\right) \tag{27}$$

for S-shaped data.

We close this section with a general theorem on the shapes of the first-citation distributions

$$\Phi(t_1) = C(t_1)^{\alpha-1}$$

as proved generally in theorem 1. In view of our result in proposition 3 the same theorem applies to all the n^th citation distributions as well.

Theorem 4 : $\Phi(t_1) = C(t_1)^{\alpha-1}$ satisfies the following properties :

a) If C is S-shaped then

a.1) if $\alpha>2$, Φ is S-shaped and the abscis of the osculation point of Φ is larger than the one of C,

a.2) if $1<\alpha<2$ and if Φ has an osculation point, then its abscis is smaller than the one of C.

b) If C is concave then

b.1) if $1<\alpha<2$ then Φ is concave,

b.2) if $\alpha>2$, Φ can be concave or S-shaped.

Of course, when $\alpha=2$, $\Phi=C$.

Proof : From $\Phi(t_1) = C(t_1)^{\alpha-1}$ it follows that

$$\Phi''(t_1) = (\alpha-1)(\alpha-2)C(t)^{\alpha-3}C'(t) + (\alpha-1)C(t)^{\alpha-2}C''(t)\,. \tag{28}$$

a.1) If C is S-shaped and $\alpha>2$ then

$$\Phi''(t_1)>(\alpha-1)C(t)^{\alpha-2}C''(t)$$

(since $C'>0$). Hence $\Phi''(t_1)>0$.
for all $t_1 \leq$ the osculation point of C (since $C''>0$ there because of the S-shape). But there exists always an osculation point of Φ since Φ strictly increases, $\Phi''>0$, and since $\lim\limits_{t_1\to+\infty} \Phi(t_1) = 1$. Hence its abscis is larger than the one of C.

a.2) If C is S-shaped and $1<\alpha<2$ then

$$\Phi''(t_1)<(\alpha-1)C(t)^{\alpha-2}C''(t).$$

Hence $\Phi''(t_1)<0$ for all $t_1 \geq$ the osculation point of C. Because of this and since we assumed the existence of the osculation point of Φ, its abscis must be smaller than the one of C.

b.1) If C is concave and $1<\alpha<2$ then it follows from (28) that $\Phi''(t_1)<0$ always. Hence Φ is concave.

b.2) If C is concave and $\alpha>2$ then Φ can be convex or S-shaped (in fact the S-shape was noted already in Egghe (1999) for the concave exponential function C of (21)). □

Note : The general philosophy behind the formula

$$\Phi(t_1) = C(t_1)^{\alpha-1}$$

is as follows. The coefficient α is well-known to be an indicator for concentration : the higher α, the less documents one has (in a relative sense) with a high number of citations (or quick first citations) and the more there are with just a few citations. High αs are considered from $\alpha>2$ on. In this case the curve of C(t) is flattened for low t by applying the power α-1. In this sense the increase of Φ is slow in the beginning. Of course, in the end, Φ goes to 1 and hence an osculation point often happens. In the opposite way, for small α, say $1<\alpha<2$ the opposite happens and we have faster increases in the beginning part of the graph of Φ. Here there are less cases where an osculation point occurs and more cases of concave increase. Of course, the shape of Φ also depends on the one of C. In this note we only discussed the influence of α on the shape of Φ.

III. Fitting first-citation data.

In this section we will try to fit four first citation data sets, using the cumulative distributions (26) and (27) :

$$\Phi(t_1) = 1-a^{t_1} \tag{29}$$

for concave data and

$$\Phi(t_1) = F\left(\frac{\ln t_1-\mu}{\sigma}\right) \tag{30}$$

for S-shaped data.

Three of the four data sets can be found in Egghe (1999). They are the Gupta and Rousseau (1999) data, the Motylev (1981) data and the Rousseau (1994) data. The fourth data set is produced by ourselves and deals with first-citation data of JASIS papers of 1980, that appear in JASIS itself in the period 1980-now (end 1999). The latter one is similar with the Rousseau data : there one examines first-citations of JACS papers by JACS papers (JACS = Journal of the American Chemical Society).

III.1 Gupta and Rousseau (1999) data

Since the cloud of points apparently is concave, we have fitted (29). We found the distribution

$$\Phi(t_1) = 1-(0.772)^{t_1}, \qquad (31)$$

hence the aging rate is a=0.772. The fit is very good (Kolmogorov-Smirnov test - see e.g. Ravichandra Rao (1983) - gives a critical value at the 5% level of $D_{0.5}$=0.338 while the maximum difference found was 0.0946). We also fitted the lognormal distribution. Although it is also acceptable, the fit was less good as the one of (31). Indeed here we found D_{max}=0.1457 (again $D_{0.5}$=0.338). The mean of log x is 0.4582 and its standard deviation is 0.6769. Of course, (31) fits less than the graph of

$$\Phi(t_1) = (1-(0.672)^{t_1})^{1.536} \qquad (32)$$

obtained in Egghe (1999) for α=2.536, but here an extra parameter (α) is involved. (31) is the simplest model that fits. See Fig. 1 for this fit as well as for the fit of the lognormal ditribution.

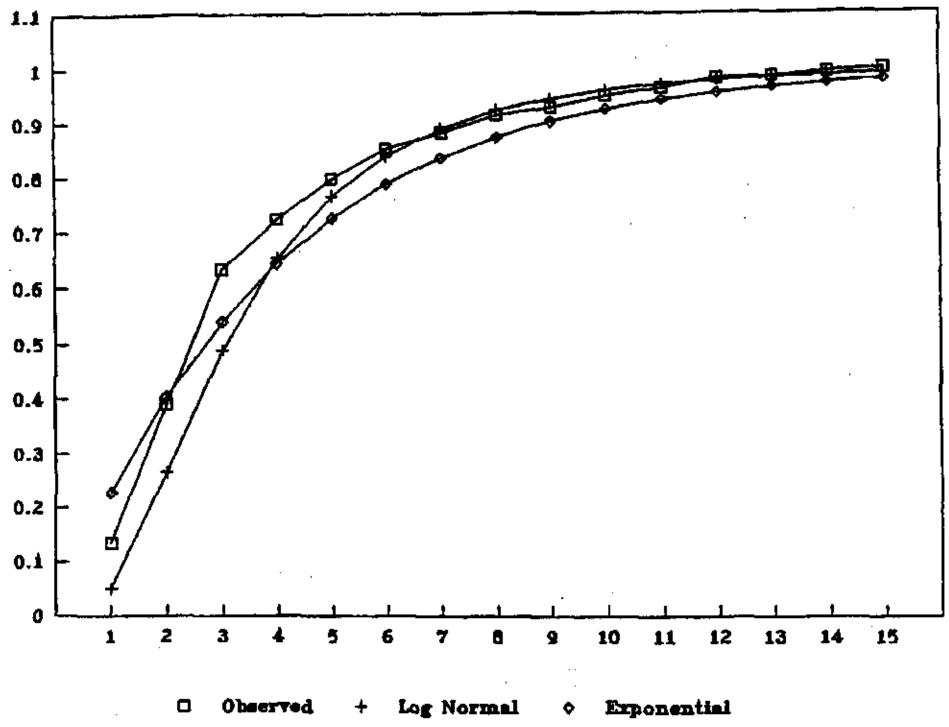

Fig. 1 Fitting the Gupta and Rousseau data

III.2 Motylev (1981) data

These data are a bit more irregular but the concave shape prevails. The exponential model (29) fits best as is also clear from visual inspection (see Fig. 2 with exponential and lognormal fit). For (29) we now have

$$\Phi(t_1) = 1-(0.863)^{t_1} \tag{33}$$

and $D_{0.5}$=0.328 while D_{max}=0.0939. For the lognormal fit we obtained D_{max}=0.1262 (again $D_{0.5}$=0.328), hence a good fit. The mean of log x is 1.612 and its standard deviation is 0.6722. In Egghe (1999) a similar fit was obtained for

$$\Phi(t_1) = (1-(0.956)^{t_1})^{0.746} \tag{34}$$

for α=1.746, involving Lotka's α, hence using a 2-parameter model.

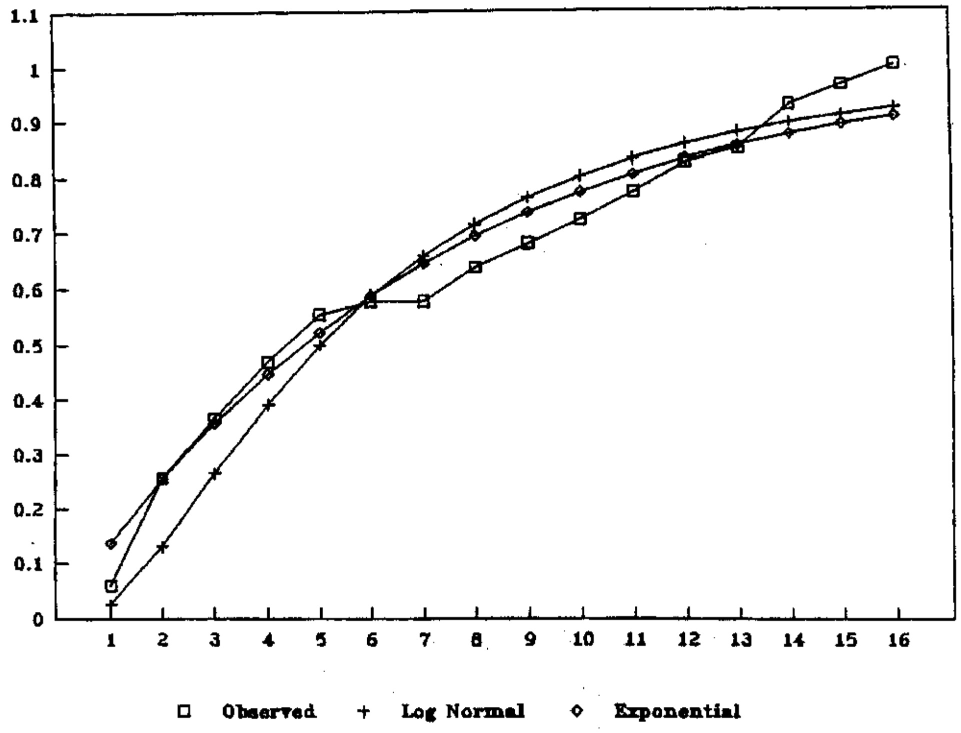

Fig. 2 Fitting the Motylev data

III.3 Rousseau (1994) data

The S-shape is clear and it is hence also clear that here the lognormal distribution (39) fits best. In fact (29) does not fit at all. For (30) we now find $D_{0.5}$=0.1347 while D_{max}=0.0453, a very good fit, which can also be seen by visual inspection (see Fig. 3). The found lognormal distribution has parameters : mean of log x is 3.3951 with standard deviation 0.7250.

In Egghe (1999) we found th distribution

$$\Phi(t_1) = (1-(0.955)^{t_1})^{2.641} \qquad (35)$$

for α=3.641. This fit is better and, since both models use 2 parameters, the latter is to be preferred.

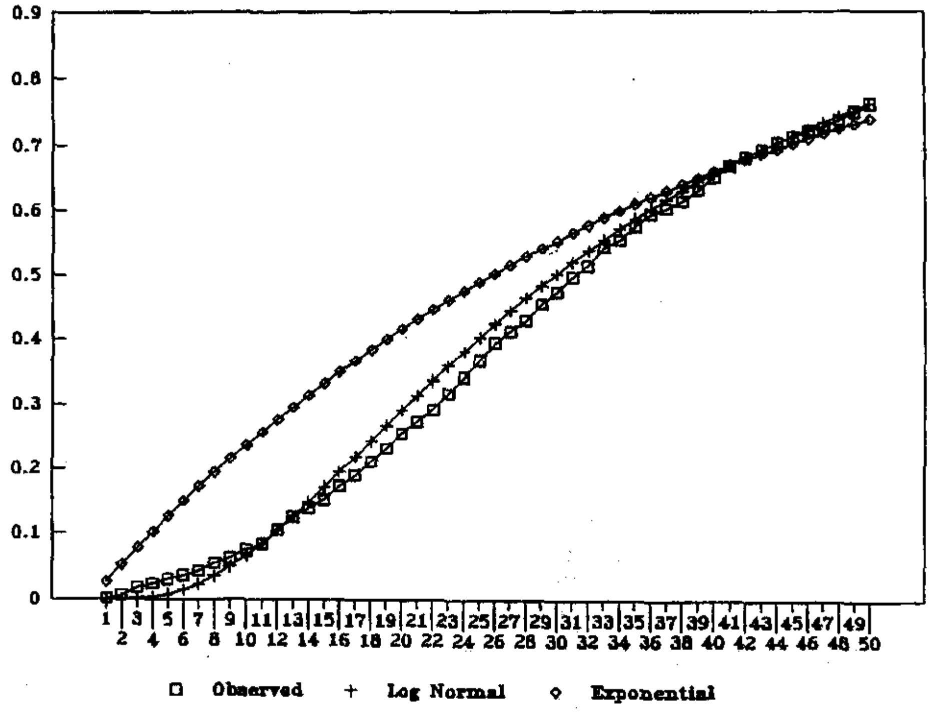

Fig. 3 Fitting the Rousseau data

III.4 New JASIS data

We examined the 1980 volume of JASIS and checked the times of first-citation in JASIS itself in the period 1980-now (end 1999). The first-citation data (cumulative fractions) are given in table I.

Table I. JASIS to JASIS data

year	cumulative fraction of first-citation
1	0.0714
2	0.3929
3	0.5714
4	0.6429
5	0.7143
6	0.7143
7	0.7500
8	0.7500
9	0.8214
10	0.8571
11	0.8929
12	0.9643
13	1.0000

It is clear from the graph that the exponential model fits best and this is also confirmed by applying Kolmogorov-Smirnov's test. The exponential model gives D_{max}=0.115 while $D_{0.5}$=0.375. See Fig. 4 for the exponential as well as the (less good) lognormal fit. Hence we have here that the simplest 1-parameter model (29) works best, giving the distribution

$$\Phi(t_1) = 1-(0.814)^{t_1} , \tag{36}$$

hence the aging rate is a=0.814.

An attempt was made to fit a lognormal distribution. We found D_{max}=0.1823 while again $D_{0.5}$=0.375, hence a good fit. The mean of log x is 1.2964 and its standard deviation is 0.7500.

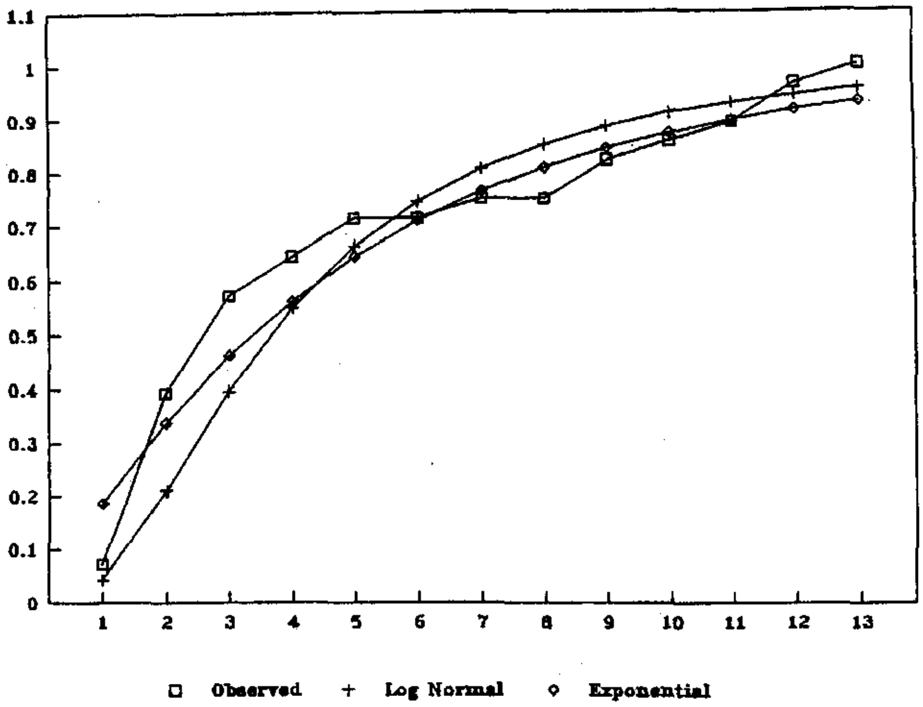

Fig. 4 Fitting the JASIS to JASIS data.

General Note : In this paper we have emphasized on finding the cumulative first-citation distribution. Hence, for $t_1 \to +\infty$ we have that $\Phi(t_1) \to 1$. Of course, here we only consider papers that eventually will be cited at least once. If we want to include the fraction γ of uncited papers it suffices to consider $\gamma\Phi$ instead of Φ, as was also done in Egghe (1999), Gupta and Rousseau (1999) and Rousseau (1994).

References

L. Egghe (1999). An informetric explanation of the first-citation distribution. Preprint.

L. Egghe and I.K. Ravichandra Rao (1992). Citation age data and the obsolescence function : fits and explanations. Information Processing and Management 28(2), 201-217.

L. Egghe and R. Rousseau (1990). Introduction to Informetrics. Quantitative Methods in Library, Documentation and Information Science. Elsevier, Amsterdam.

B.M. Gupta and R. Rousseau (1999). Further investigations into the first-citation process : the case of population genetics. Libres, to appear.

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

E. Matricciani (1991). The probability distribution of the age of references in engineering papers. IEEE Transactions of Professional Communication 34, 7-12.

V.M. Motylev (1981). Study into the stochastic process of change in the literature citation pattern and possible approaches to literature obsolescence estimation. International Forum on Information and Documentation 6, 3-12.

I.K. Ravichandra Rao (1983). Quantitative Methods to Library and Information Science. Wiley Eastern, New Delhi, 1983.

R. Rousseau (1994). Double exponential models for first-citation processes. Scientometrics 30(1), 213-227.