PERGAMON

Mathematical and Computer Modelling 34 (2001) 81–90

MATHEMATICAL AND COMPUTER MODELLING

www elsevier nl/locate/mcm

Theory of First-Citation Distributions and Applications

L. EGGHE
LUC, Universitaire Campus, B-3590 Diepenbeek, Belgium†
and
DRTC, ISI, 8th Mile Mysore Road
R V College P O , Bangalore 560059, India‡
and
UIA, Universiteitsplein 1, B-2610 Wilrijk, Belgium
leo egghe@luc ac be

I K RAVICHANDRA RAO
DRTC, ISI, 8th Mile Mysore Road
R V College P O , Bangalore 560059, India
ikr@isibang ac in

(Received May 2000, accepted June 2000)

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 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 $n \in \mathbb{N}$ © 2001 Elsevier Science Ltd All rights reserved

Keywords—First-citation distribution, Lotka, Exponential distribution, Lognormal distribution

1. 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 reuse the few first-citation

†Permanent address
‡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

PII S0895-7177(01)00050-4

data that are available data of Gupta and Rousseau [1] in theoretical population genetics, the Motylev data appearing in [2] on Russian language library science papers, and the JACS (*Journal of the American Chemical Society*) data of Rousseau [3] 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

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

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 [4,5]) 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

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

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 $c = -\ln a$), or

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

the lognormal distribution (t continuous)

The exponential distribution is the most appealing and is basic to all citation studies (see, e g , [5]) 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 logarithmic time, $\ln t$ The lognormal distribution has been generally accepted and explained, see e g , [6,7]

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 $\alpha = 2$ (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 ($\alpha = 2$), 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 data [1], the Motylev data [2], the Rousseau data [3], and our JASIS data are fitted very well

Of course, as noted in [8], 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 [8] to fit Rousseau's JACS data as well as the Gupta and Rousseau data (although the S-shape is not very 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 three parameters (μ, σ, α) and fitting powers of cumulative lognormal distributions are not part of standard statistical packages All our fittings involve only two parameters

2. 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 conditional with respect to the ever cited documents The fraction γ of the noncited 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 was 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

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

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

$$\Phi(t_1) = C(t_1)^{\alpha-1}$$ ∎

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 ($n \in \mathbb{N}$ 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 $n \in \mathbb{N}$, 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 $n \in \mathbb{N}$ The fraction (among 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 (among 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

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

∎

So again, if $\alpha = 2$, we see that $\Phi_n = C$ for all $n \in \mathbb{N}$

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 $c = -\ln a$ The cumulative distribution then takes the form

$$C(t) = \int_0^t -(\ln a)a^{t'}\,dt', \qquad C(t) = 1 - a^t \tag{21}$$

In this case, Theorem 1 says that

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

hence, we again find the result in [8] Note, that in [8] 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 [8], distribution (22) has proved to fit very well first-citation data such as the ones of [1–3] Distribution (22) is capable of fitting concave as well as S-shaped data Indeed, as proved in [8], (22) is concave iff $1 < \alpha \le 2$ 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 in this in more detail If

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

its cumulative distribution is

$$\begin{aligned} C(t) &= \int_0^t \frac{1}{\sqrt{2\pi}\sigma t}e^{-(1/2)((\ln t-\mu)/\sigma)^2}\,dt = \int_{-\infty}^{(\ln t-\mu)/\sigma} \frac{1}{\sqrt{2\pi}}e^{-(1/2)s^2}\,ds, \\ C(t) &= F\left(\frac{\ln t-\mu}{\sigma}\right), \end{aligned} \tag{24}$$

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

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

foi 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*

  (a1) *if $\alpha > 2$, Φ is S-shaped and the abscissa of the osculation point of Φ is larger than the one of C,*

  (a2) *if $1 < \alpha < 2$ and if Φ has an osculation point, then its abscissa is smaller than the one of C*

(b) *If C is concave, then*

  (b1) *if $1 < \alpha < 2$, then Φ is concave,*

  (b2) *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

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

(a1) 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 $\lim_{t_1 \to +\infty} \Phi(t_1) = 1$ Hence, its abscissa is larger than the one of C

(a2) 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 abscissa must be smaller than the one of C

(b1) If C is concave and $1 < \alpha < 2$, then it follows from (28) that $\Phi''(t_1) < 0$ always Hence, Φ is concave

(b2) If C is concave and $\alpha > 2$, then Φ can be convex or S-shaped (in fact, the S-shape was noted already in [8] 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 fewer 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 $\alpha - 1$ 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 fewer 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

3. 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 [8] They are the Gupta and Rousseau data [1], the Motylev data [2] and the Rousseau data [3] 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*)

3.1. Gupta and Rousseau Data [1]

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

$$\Phi(t_1) = 1 - (0\ 772)^{t_1}, \tag{31}$$

hence, the aging rate is $a = 0\ 772$ The fit is very good (Kolmogorov-Smirnov test—see e g , [9]—gives a critical value at the 5% level of $D_{0\ 5} = 0\ 338$ while the maximum difference was 0 0946) We also fitted the lognormal distribution Although it is also acceptable, the fit was not as 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

$$\Phi(t_1) = \left(1 - (0\ 672)^{t_1}\right)^{1\ 536}, \tag{32}$$

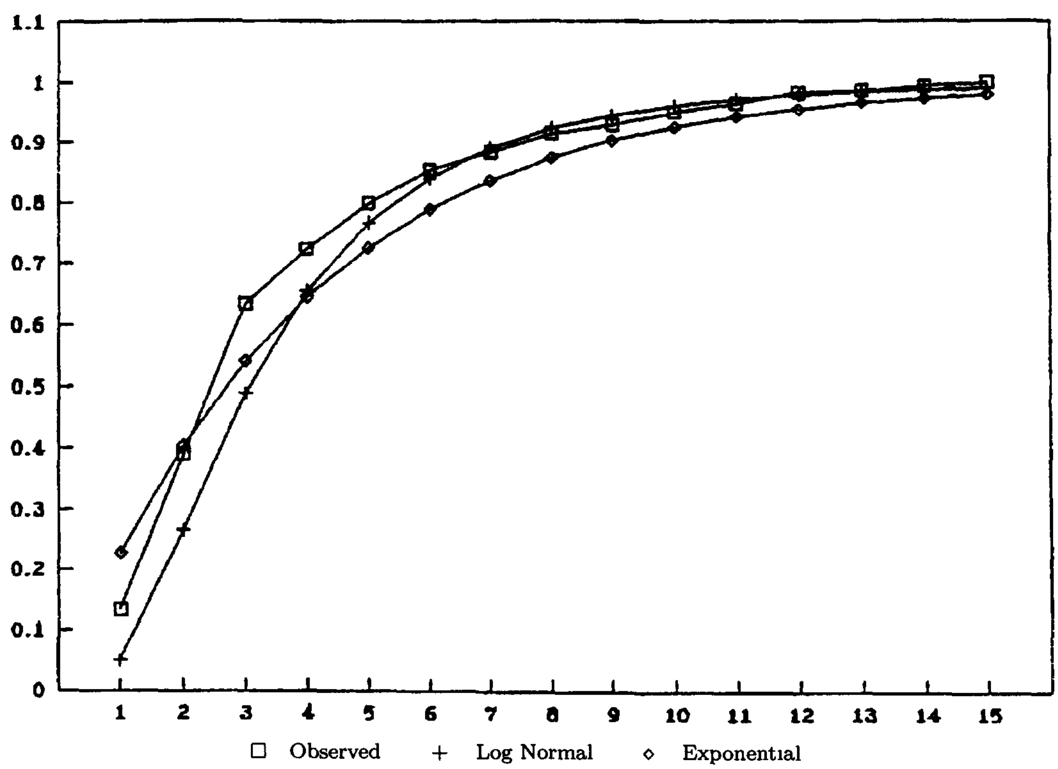

Figure 1 Fitting the Gupta and Rousseau data

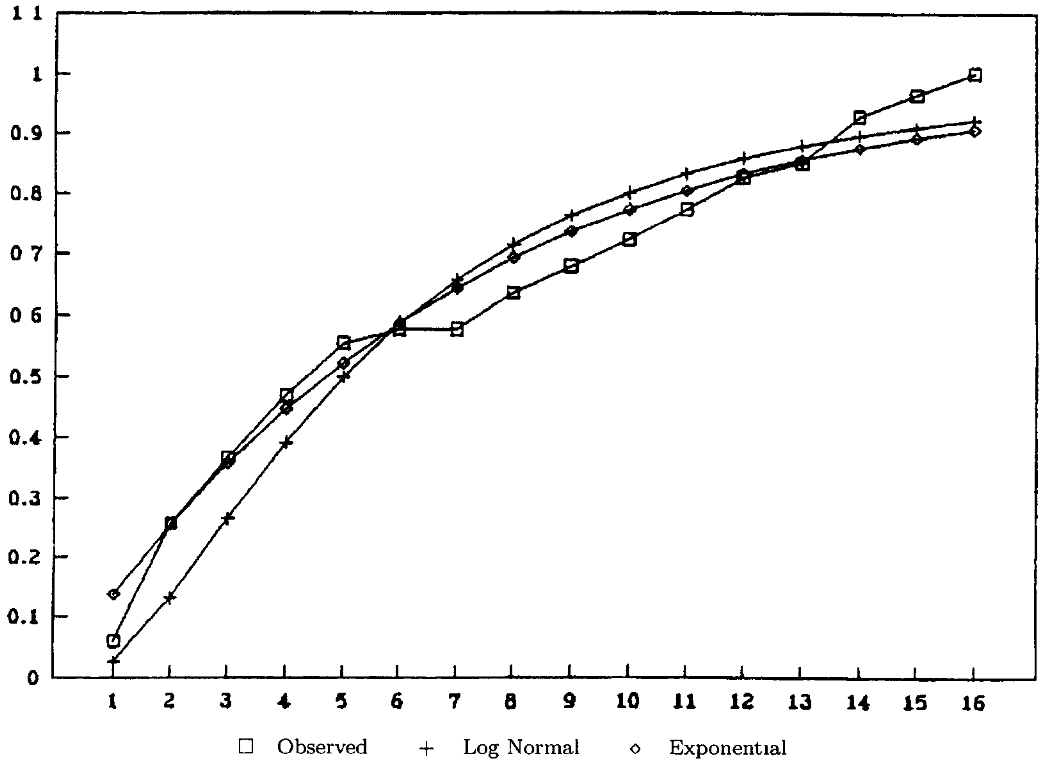

Figure 2 Fitting the Motylev data

obtained in [8] for $\alpha = 2\ 536$, but here an extra parameter (α) is involved Equation (31) is the simplest model that fits See Figure 1 for this fit, as well as for the fit of the lognormal distribution

3.2. Motylev Data [2]

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 Figure 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 [8], a similar fit was obtained for

$$\Phi(t_1) = \left(1 - (0\ 956)^{t_1}\right)^{0\ 746}, \tag{34}$$

for $\alpha = 1\ 746$, involving Lotka's α, hence using a 2-parameter model

3.3. Rousseau Data [3]

The S-shape is clear and it is hence also clear that here the lognormal distribution (30) 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 Figure 3) The found lognormal distribution has parameters mean of $\log x$ is 3 3951 with standard deviation 0 7250

In [8], we found the distribution

$$\Phi(t_1) = \left(1 - (0\ 955)^{t_1}\right)^{2\ 641}, \tag{35}$$

for $\alpha = 3\ 641$ This fit is better and, since both models use two parameters, the latter is to be preferred

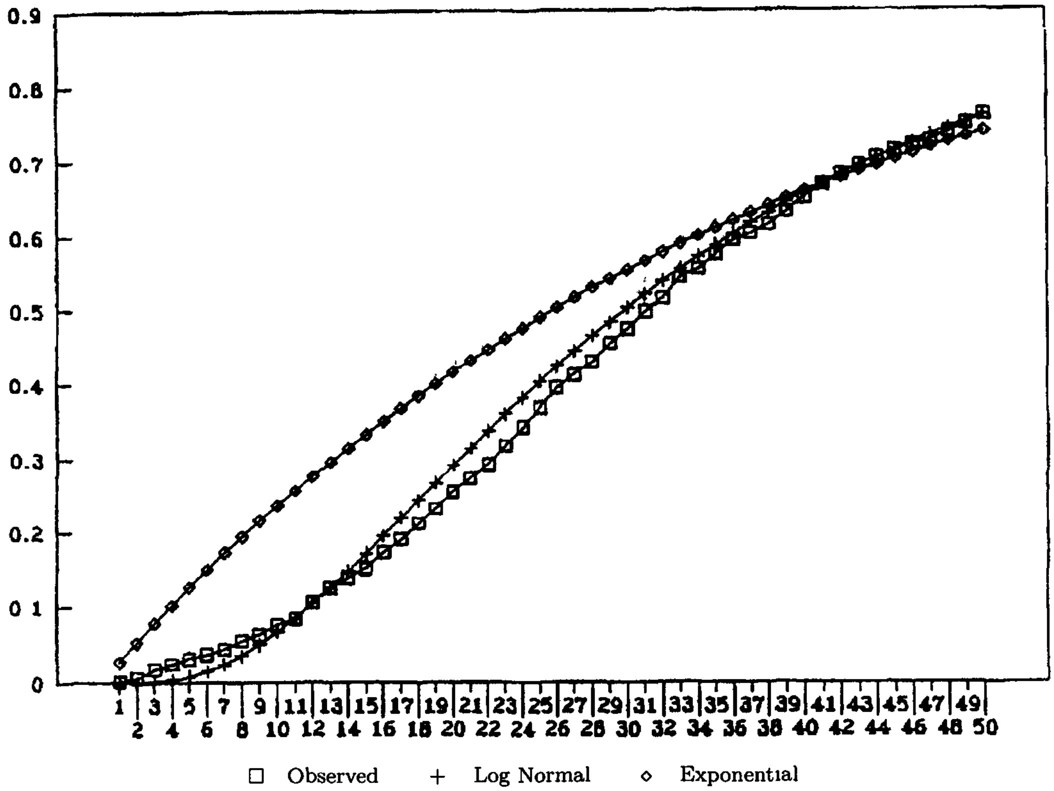

Figure 3 Fitting the Rousseau data

3.4. 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 1

Table 1 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 Figure 4 for the exponential as well as the (not as good) lognormal fit Hence,

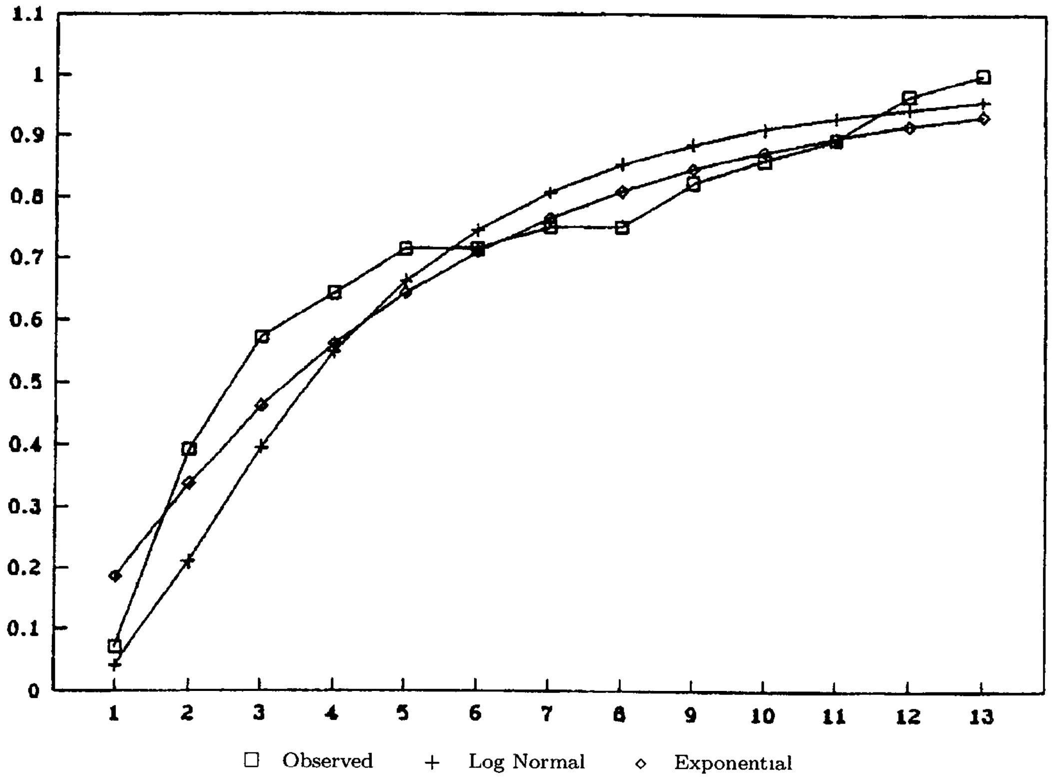

Figure 4 Fitting the JASIS to JASIS data

we have here that the simplest one-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

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, if suffices to consider $\gamma\Phi$ instead of Φ, as was also done in [1,3,8]

REFERENCES

1 B M Gupta and R Rousseau, Further investigations into the first-citation process The case of population genetics, *Libres* **2** (1999), `http //aztec lib utk edu/libres/libregn2/fc ftm`

2 V M Motylev, 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 (1981)

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

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

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

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

7 L Egghe and I K Ravichandra Rao, Citation age data and the obsolescence function Fits and explanations, *Information Processing and Management* **28** (2), 201–217 (1992)

8 L Egghe, A heuristic study of the first-citation distribution, *Scientometrics* **48** (3), 345–359 (2000)

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