The source-item coverage of the exponential function

Thierry Lafouge

The source-item coverage of the exponential function

Thierry Lafouge
Laboratoire Elico Université Claude Bernard Lyon1
43 Boulevard du 11 novembre 1918 69622 Villeurbanne Cedex, France
Lafouge@univ-lyon1.fr

I dedicate this work to the memory of B. Delattre, brilliant mathematician.

Abstract

Statistical distributions in the production of information are most often studied in the framework of Lotkaian informetrics. In this article, we recall some results of basic theory of Lotkaian informetrics, then we transpose methods (Theorem1) applied to Lotkaian distributions by Leo Egghe (Theorem 2) to the exponential distributions (Theorem 3, Theorem 4). We give examples and compare the results (Theorem 5). Finally, we propose to widen the problem using the concept of exponential informetric process (Theorem 6).

Keywords: Exponential function; mathematical-fitting; Lotkaian informetrics.

1. Introduction

Many phenomena studied in informetrics, concerning the production or use of information, can be represented by a triple (Source, Production function, Items) called Information Production Process (IPP) (Egghe, 1990). This consists of a set of sources S, a set of items I, T (respectively A) denotes the total number of sources, respectively the total number of items, and finally a function of production or use that quantifies the production of the items by the sources. There are several methods for representing these phenomena. In this article, the theory is developed with a size-frequency function f, the most usual form for quantifying this production.

$f : [1, I_{\max}] \to \Re^+$, $f(j)$ describes the density of sources with item density j (in the discrete setting $f(j)$ indicates the number of sources that have produced j items). We assume f is continuous.
$I_{\max}$ indicates the maximal item per source density.
The following two equalities allow us to calculate T and A.

$$T = \int_1^{I_{\max}} f(j)dj \tag{1.1}$$

$$A = \int_1^{I_{\max}} j.f(j)dj \tag{1.2}$$

Of course it may happen that T and A are infinite if $I_{\max}$ is infinite. If T is finite we have the following inequality: $0 < T < A$ .We denote $\mu = \frac{A}{T}$, the average number of items by source. We have, $\mu > 1$.

In practice, the production function of an IPP has similar characteristics in very diverse situations of production or use of information:

- Authors (Sources) write articles (Items)
- Words in a text (Sources) produce occurrences of words in the text (Items)
- Web pages (Sources) contain links (Items)
- Web sites (Sources) are visited (Items)
- Requests (Sources) through a search engine are sent by users (Items).

In all quoted examples, if we quantify the production of the items by the sources with a size-frequency function, this one is decreasing with a long tail and a gap between a high number of sources producing few items and a small number of sources producing a lot. In practice, when one determines a best-fitting curve, we must truncate the distribution because for high frequencies the number of items produced is very low. This characteristic results in the standard deviation often being extremely high compared with the average and is a poor indicator. The statistical distribution most used in informetrics is the inverse power function, also called Lotkaian informetric distribution. This distribution is unimodal; it models the information production processes in many of the quoted examples . At present, with Internet, there are many examples for which the data resulting from the Web has been adjusted by such models (Bilke & Peterson, 2001). The GIGP (Zero-truncated Generalized Inverse Gaussian-Poisson) model known and tested over a long time (Burrel & Fenton, 1993) is also used to day to adjust some of this data (Ajiferuke & Wolfram, 2004).

2. Lotkaian informetric distribution

With the preceding notations, a Lotkaian informetric distribution is given:

$$f : [1, I_{\max}] \to \Re^{+},$$

where

$$f(j) = \frac{C}{j^{\alpha}}, C > 0 \text{ and } \alpha > 1 \qquad (2.1)$$

We will limit ourselves to the case where $\alpha > 1$, meaning where T (number of sources) is finite and where the corresponding probability density function is: $f(j) = (\alpha - 1).j^{-\alpha}$ $\alpha > 1$, if $I_{\max} = \infty$. Moreover, we know that A is finite if $\alpha > 2$. More generally f has moments of order n if $\alpha > n$.

The mathematical properties of these functions (Haitun, 1982) have been to great extent studied. They have often been opposed to the functions modeling Gaussian processes. They have been the subject of a recent work of informetrics (Egghe, 2005), which contains many results. This work has the merit among others of unifying all the work done concerning empirical applications, Lotka, Bradford, Zipf, Mandelbrot, with the mathematical theory of IPP, choosing, as central distribution, the Lotkaian distributions. The coefficient α characterizes the gap between strongly productive sources and those that produce little. Many works (Bookstein, 1990a, Bookstein, 1990b) have shown the strength of the law of Lotka (Lotka, 1926). In addition, the value $\alpha = 2$ plays a key role since we know, according to whether α is smaller or greater than 2, that the representation of Leimkuhler (Rousseau, 1988) has or does not have a turning point; in informetrics the term “Groos droop” (Groos,1967) is often used.

Finally, these distributions are scale free. A function f is called scale-free if, for every positive constant *C*, there is a positive constant *D* such that $f(Cx) = Df(x)$ for all x in the domain of f (Egghe, 2005 p 27). This property is important when frequencies are observed. It allows us to change scale without changing model. In the main, it justifies the choice of Leo Egghe in his work that we have just quoted. If we impose the scale free property, it implies that some decreasing functions, such as the decreasing exponential function one, are not allowed.
However, the phenomena of obsolescence and growth of quotations on a subject of search in scientific literature (Egghe, 1993) are modeled by exponential processes.
In addition, the often ignored result of Naranan (Naranan, 1971), shows that distributions of Lotkaian type can be deduced under certain conditions from an exponential growth of sources and the number of items produced by these sources. It gives the exponential functions an importance that we cannot neglect. In the previously quoted examples, the temporal parameter plays the role of variable.
Finally the law of geometrical probability, which is the discrete version of the exponential law, is often used as a rough approximation for modeling the processes of commands or library circulation data (Bagust, 1983).
Thus, all these reasons lead us to adopt a procedure similar to that of Leo Egghe (Egghe, 2005) and to find a mathematical result for the exponential distributions, which is a necessary condition of the same type as Lotkaian distributions.

3. Reminder of some results of basic theory of Lotkaian informetrics

There is a lot of statistical work that consists of verifying the statistical regularities in the variety of examples mentioned in the introduction, and making fittings. Lotkaian distributions play an important role here. However, to my knowledge, few bibliometric researchers use the mathematical theorem, certainly recent, which we remind is a necessary condition for the production of sources covering the items produced. This theorem plays a key role in this article.

Theorem 1 (Egghe 2005 p 111)
The following assertions are equivalent, given $A > T > 0$

(i) There exists a function $f : [1,+\infty[\to \Re^+$ and a finite number $I_{\max} > 1$ such that:

$$T = \int_1^{I_{\max}} f(j)\,dj$$

$$A = \int_1^{I_{\max}} j.f(j)\,dj$$

(ii) There exists a function $f^* : [1,+\infty[\to \Re^+$ such that

$$\mu < \frac{\int_1^{\infty} j.f^*(j)\,dj}{\int_1^{\infty} f^*(j)\,dj}$$

Moreover if (i) or (ii) holds we have, necessarily that $f^* = D.f$ with $D > 0$, a constant.

We refer readers interested in the demonstration to the reference quoted. What is interesting for us here in the theorem is the implication $(ii) \Rightarrow (i)$.

Leo Egghe applies this theorem to Lotkaian informetric distributions.

Theorem 2 (Egghe 2004)

Let $0 < T < A < \infty$ be given. Let $\alpha > 1$ and a number $I_{\max} > 1$.

If $I_{\max}$ is infinite

(i) If the inverse power function as in (2.1) satisfies (1.1) and (1.2) if we have

$$\alpha = \frac{2\mu - 1}{\mu - 1} \tag{3.1}$$

and

$$C = \frac{A}{\mu - 1} \tag{3.2}$$

which implies $\alpha > 2$ if $A < \infty$.

If $I_{\max}$ is finite (the general case)

(ii) If $\alpha \leq 2$ then there always exists a number $I_{\max} > 1$ such that (2.1) satisfies (1.1) and (1.2).

(iii) If $\alpha > 2$ the conclusion (i) is valid if and only if

$$\mu < \frac{\alpha - 1}{\alpha - 2} \tag{3.3}$$

We will follow exactly the same procedure for the exponential functions (Theorem 3 and Theorem 4), then compare the results (Theorem 5).

4. Exponential distribution

4.1 Theoretical results

With the preceding notations, an exponential function : $g : [1, I_{\max}] \to \Re^+$ is given where

$$g(j) = C.e^{-\alpha(j-1)} \text{ with } C > 0 \text{ and } \alpha > 0 \tag{4.1}$$

We can also write it in an equivalent form:

$$g(j) = C.a^{-j} \text{ with } C > 0 \text{ and } a > 1$$

We will use the first form here. The corresponding probability density function is :

$$g(j) = \alpha.e^{-\alpha.(j-1)} \text{ with } \alpha > 0 \text{ and } I_{\max} = \infty$$

As for the power function, g has as a maximum value for 1.

Unlike the inverse power function, a decreasing exponential function has moments of order n whatever the n positive.

Lemma

If we call $A(n) = \int_1^{\infty} j^n . e^{-\alpha(j-1)} dj$ the moment of order n divided by α of an exponential function,

we have $A(n) = \sum_{p=0}^{p=n-1} \frac{n!}{(n-p)!} . \frac{1}{\alpha^{p+1}} + n! . \frac{1}{\alpha^{n+1}}, \quad n \geq 0$;

Proof

An integration by part gives: $A(n) = \frac{1}{\alpha} + \frac{n}{\alpha} . A(n-1)$ with $A(0) = \frac{1}{\alpha}$ .We show by

recurrence: $A(n+1) = \frac{1}{\alpha} + \frac{(n+1)}{\alpha} . A(n) = \frac{1}{\alpha} + \frac{(n+1)}{\alpha} (\sum_{p=0}^{p=n-1} \frac{n!}{(n-p)!} . \frac{1}{\alpha^{p+1}} + n! . \frac{1}{\alpha^{n+1}})$

$= \frac{1}{\alpha} + (\sum_{p=0}^{p=n-1} \frac{(n+1)!}{(n-p)!} . \frac{1}{\alpha^{p+2}} + (n+1)! . \frac{1}{\alpha^{n+1}}) = \sum_{p=0}^{p=n} \frac{(n+1)!}{(n+1-p)!} . \frac{1}{\alpha^{p+1}} + (n+1)! . \frac{1}{\alpha^{n+2}}$

More particularly we obtain:

$$A(1) = \frac{1}{\alpha} + \frac{1}{\alpha^2} \tag{4.2}$$

□

Problem

What are the conditions, given A and T $(0 < T < A)$, for the existence of an exponential function g as in (4.1) which verifies: $\int_1^{I_{max}} g(j) dj = T$ *and* $\int_1^{I_{max}} j . g(j) dj = A$?

We separate the study into two cases.

1) $I_{\max}$ is infinite

Theorem 3

Let $0 < T < A$ be given. The exponential function defined in (4.1) satisfies the following conditions:

$$\int_1^{\infty} g(j) dj = T \; ; \int_1^{\infty} j . g(j) dj = A$$

if

$$\alpha = \frac{T}{A-T} = \frac{1}{\mu - 1} \tag{4.3}$$

$$C = \frac{T^2}{A-T} \tag{4.4}$$

Proof

By solving the integrals above (see (4.2)), we obtain:

$$\int_1^{\infty} C.e^{-\alpha(j-1)}dj = \frac{C}{\alpha}$$

and

$$\int_1^{\infty} C.j.e^{-\alpha(j-1)}dj = C\left(\frac{1}{\alpha^2}+\frac{1}{\alpha}\right)$$

.

We then deduce the desired results ((4.3), (4.4)) by solving the two equations:

$$T = \frac{C}{\alpha}$$

and

$$A = \frac{C.\alpha + C}{\alpha^2}.$$

As for a Lotkaian distribution, α only depends on μ. When fitting, the formulas ((4.3),(4.4) give a rough estimate of the parameters of the exponential function (4.1).

□

2) $I_{\max}$ is finite

Theorem 4

Let $A > T > 0$ be given. Thus $\alpha > 0$, there is still $I_{\max} > 1$ finite and an exponential function defined by (4.1) verifying the two conditions:

$$\int_1^{I_{\max}} g(j)dj = T \quad and \quad \int_1^{I_{max}} j.g(j)dj = A$$

if the inequality

$$\mu < 1 + \frac{1}{\alpha} \tag{4.5}$$

holds.

Proof

According to the preceding lemma, we have : $\dfrac{\int_1^{\infty} j.Ce^{-\alpha(j-1)}dj}{\int_1^{\infty} Ce^{-\alpha(j-1)}dj} = \dfrac{\frac{1}{\alpha^2}+\frac{1}{\alpha}}{\frac{1}{\alpha}} = 1+\frac{1}{\alpha}$

The assertion (ii) of theorem 1 allows us to conclude. It will be noticed that the result is valid for any value $\alpha > 0$. In particular, the value $\alpha = 2$, unlike the Lotkaian distribution, is not a key value.

Construction of g

Now its existence is proven, we must show how to build it. To simplify the notations we put $x = I_{\max}$.

$$\int_1^x Ce^{-\alpha(j-1)}dj = T \quad \Rightarrow \quad \frac{T}{C} = \frac{1-e^{-\alpha(x-1)}}{\alpha}$$

$$\int_1^x j.Ce^{-\alpha(j-1)} = A \quad \Rightarrow \quad \frac{A}{C} = \frac{1-x.e^{-\alpha(x-1)}}{\alpha} + \frac{1-e^{-\alpha(x-1)}}{\alpha^2}$$

We suppose $x \neq 1$. By eliminating C we deduce the following equation:

$$\frac{A.\alpha}{T} = \frac{e^{-\alpha(x-1)}.(-1-x.\alpha)+\alpha+1}{(1-e^{-\alpha(x-1)}))}$$

thus

$$\mu\alpha - \frac{e^{-\alpha(x-1)}.(-1-x.\alpha)+\alpha+1}{(1-e^{-\alpha(x-1)})} = 0 \qquad (4.6)$$

Unlike the case where x is infinite there are many values $\alpha > 0$ where the preceding equation has solutions. We consider α as a parameter of the equation (4.6), $\alpha > 0$. We solve this equation in x, by the iterative method, using the MAPPLE 4.0 software for example. Then we calculate C,

$$C = \frac{\alpha.T}{1-e^{-\alpha(x-1)}} \qquad (4.7)$$

□

We have just seen a necessary condition for an exponential function to produce a given number of items with a given number of sources. We shall see that if this necessary condition holds, then it also holds for a Lotkaian distribution. More precisely, we have the following result:

Theorem 5

Let $A > T > 0$ be given. Let $\alpha > 1$. If there is a number $I_{\max} > 1$ such that (4.1) satisfies (1.1) and (1.2), then it is also valid for (2.1).

Proof

If $I_{\max} = \infty$ it is still true according to the results (i) of theorem 2 .

If $I_{\max}$ is finite we have two cases.

(i) $\alpha \leq 2$, we know according to the result (ii) of theorem 2 that it is also true.

(ii) $\alpha > 2$, we know according to theorem 4 that (4.5) is true, thus $\alpha < \frac{1}{\mu - 1}$, then $\frac{1}{\mu-1} < \frac{2.\mu-1}{\mu-1}$ thus the inequality $\alpha < \frac{2\mu-1}{\mu-1}$ is true, thus the inequality (3.3) is true: , the assertion (iii) of theorem 2 then allows us to conclude.

□

4.2 Examples

1) $A = 10,000$, $T = 5,000$ thus $\mu = 2$ and $\alpha = 0.5$. The inequality (4.5) is demonstrated.

We must then solve the equation (4.6) : $1 - \frac{e^{-0..5(x-1)}.(-1-0.5.x)+1.5}{1-e^{-0.5(x-1)}} = 0$. We obtain the solution $x = 3.512$, then according to (4.7) we have $C \approx 8,778$

The desired exponential function is: $g(j) = 8,778\, e^{-0,5.(j-1)}$

2) $A = 10,000$, $T = 7,000$ thus $\mu = 1.43$ and $\alpha = 2$. The inequality (4.5) is demonstrated.

We must then solve the equation (4.6) : $2.86 - \frac{e^{-2(x-1)}.(-1-2.x)+3}{1-e^{-2(x-1)}} = 0$. We obtain the solution $x = 2.58$ then according to (4.7) we have $C \approx 14,62$

The desired exponential function is: $g(j) = 14,62\, e^{-2.(j-1)}$.

Note

The results in section 4 could be considered as a "mathematical fitting" method for exponential function, as opposed to statistical fitting.

5. Perspectives: exponential informetric process

In the article (Lafouge & Prime-Claverie, 2005) we define an exponential informetric process in terms of an exponential function and an effort function where the average quantity supplied by the sources, to produce all the items is finite. More precisely, a set of functions, denoted EF.

EF = $\{ h : [1,\infty[\to \Re^{+}$: increasing, continuous, and not majorized $\}$

$h \in$ EF an effort function is then any element of EF.

, we call exponential informetric process the size-frequency function $\nu(h)$:

$$\nu(h)(j) = C.e^{-h((j)} \quad C > 0 \tag{5.1}$$

where the following quantity

$$F = \int_{1}^{\infty} \nu(h)(j).h(j)dj \tag{5.2}$$

is finite, F corresponds to the quantity of effort produced by $\nu(h)$.

Note

The total number of sources *T*, $T = \int_{1}^{\infty} \nu(h)(j).dj$ is finite and we have the inequality $\infty > F > T > 0$.

Examples

The respective functions of effort, $h(j) = \alpha.Ln(j), \alpha > 1$ and $h(j) = \alpha(j-1), \alpha > 0$ correspond to the inverse power function $f(j) = \frac{C}{j^{\alpha}}$ and to the exponential function $g(j) = C.e^{-\alpha(j-1)}$, studied previously.

Problem

What are the conditions, given the quantity of effort F, the number of sources T, for the existence of an exponential informetric process $\nu(h)$ as in (5.1), where $I_{\max}$ is a number >1, which verifies $F = \int_1^{i_{\max}} \nu(h)(j).h(j)dj$ and $T = \int_1^{i_{\max}} \nu(h)(j).dj$?

We will limit ourselves to the case where the respective functions of effort correspond to the inverse power function (2.1) and to the exponential function (4.1), and where $I_{\max} = \infty$.

Theorem 6

Let $F > T > 0$ be given:

(i) the exponential informetric process as in (5.1) where $h(j) = \alpha(j-1), \alpha > 0$ satisfies the following conditions:

$$T = \int_1^{\infty} C.e^{-\alpha(j-1)}.dj$$

$$F = \int_1^{\infty} C.e^{-\alpha(j-1)}.\alpha(j-1).dj$$

if

$$T = F = \frac{C}{\alpha} \qquad (5.3)$$

(ii) the exponential informetric process as in (5.1) where $h(j) = \alpha.Ln(j), \alpha > 1$ satisfies the following conditions:

$$T = \int_1^{\infty} C.\frac{1}{j^{\alpha}}.dj$$

$$F = \int_1^{\infty} C.\frac{1}{j^{\alpha}}\alpha.Ln(j).dj$$

if

$$\alpha = \frac{F}{F - T} \qquad (5.4)$$

$$C = \frac{T^2}{F - T} \tag{5.5}$$

Proof

(1) By (4.1) $T = \frac{C}{\alpha}$. An integration by part give: $F = \frac{C}{\alpha}$

(2) By (2.1) $T = \frac{C}{\alpha - 1}$

$\int_1^\infty \frac{1}{j^\alpha}.Ln(j).dj = -\frac{1}{\alpha - 1}\int_1^\infty Ln(j)d(\frac{1}{j^{\alpha-1}})$, an integration by part give : $\int_1^\infty \frac{1}{j^\alpha}.Ln(j).dj = \frac{1}{(\alpha - 1)^2}$ thus $F = \frac{C.\alpha}{(\alpha - 1)^2}$

We then deduce the desired results (5.4) and (5.5) by solving the two equations:

$$T = \frac{C}{\alpha - 1}$$

and

$$F = \frac{C.\alpha}{(\alpha - 1)^2}$$

□

The case where $I_{\max}$ is finite is an open problem.

References

Ajiferuke, I. & Wolfram, D. (2004). Informetric modelling of Internet Search and Browsing Characteristics. *The Canadian Journal of information and library Science*, 28(1), 1-16.

Bagust, A. (1983). A circulation model for busy public libraries. *Journal of documentation*, 39(1), 24-37.

Bilke, S. & Peterson, C. (2001). Topological properties and metabolic networks. *Physical Reviews E*. 6403(3), 76-80.

Bookstein, A.(1990a). Informetric Distribution, Part 1: Unified Overview. *Journal of the American Society for Information Science*, 41(5), 368-375.

Bookstein, A.(1990b). Informetric Distribution, Part 2: Resilience to Ambiguity. *Journal of the American Society for Information Science*, 41(5), 376-385.

Burrel,Q.L & 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.

Egghe, L. (1990). On the duality of informetric systems with applications to the empirical law. *Journal of Information Science*, 16, 17-27.

Egghe, L. (1993). On the influence of growth on obsolescence. *Scientometrics* 27 (1), 195-214.

Egghe, L. (2004). The source-item coverage of the Lotka function. *Scientometrics*, 61(1), 103-115.

Egghe, L. (2005). Power laws in the Information Production Process: Lotkaian Informetrics. Elsevier.

Groos, AV. (1967). Bradford's law and the Keenan-Atherton data. *American Documentation*,18, 46.

Haitun, S.D.(1982) Stationary Scientometric Distributions. *Scientometrics* n°4, 1982, Part I,5-25, Part II, 89-104, Part III, 181-194.

Lafouge, T. & Prime Claverie, C. (2005). Production and use of information. Characterization of informetric distributions using effort function and density function Exponential informetric process. *Information Processing and Management*, 41, 1387-1394.

Lotka, A.J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Science*, 292-306.

Naranan, S. (1971). Bradford Law of bibliography of science an interpretation. *Nature*, 227(5258), 631-632.

Rousseau, R. (1988). Lotka's law and its Leimkuhler representation. *Library Science with a Slant to Documentation and Information studies*, 25(3), 150-178.