A characterization of the law of Lotka in terms of sampling

EGGHE, Leo (2005) A characterization of the law of Lotka in terms of sampling. In: Scientometrics, 62(3). p. 321-328.

DOI: 10.1007/s11192-005-0024-6
Handle: http://hdl.handle.net/1942/740

A characterization of the law of Lotka in terms of sampling

by

L. Egghe

Limburgs Universitair Centrum (LUC), Universitaire Campus, B-3590 Diepenbeek, Belgium[1]
and
Universiteit Antwerpen (UA), Campus Drie Eiken, Universiteitsplein 1, B-2610 Wilrijk, Belgium

leo.egghe@luc.ac.be

ABSTRACT

An incomplete bibliography (or, more generally, an incomplete Information Production Process (IPP)) can be considered as a sample from a complete one. Sampling can be done in the sources or in the items. The simplest sampling technique is the systematic one where every k^{th} source or k^{th} item is taken (alternatively: deleted) ($k \in \mathbb{N}$).

In this paper we give a definition of systematic sampling in items and sources in the framework of an IPP in which we have continuous variables. We prove the theorem that in such IPPs we have a Lotkaian size-frequency function (i.e. a decreasing power function) if and only if systematic sampling in sources is the same as systematic sampling in items. In this proof we use the well-known characterization of power functions as scale-free functions.

---

[1] Permanent address: Limburgs Universitair Centrum, Universitaire Campus, B-3590 Diepenbeek, Belgium.
Key words and phrases: Lotka, systematic sample, source, item, scale-free.

I. Introduction

In this paper we will use the terminology "Information Production Process" (IPP) for a generalized bibliography, where one has a set of sources (e.g. journals), a set of items (e.g. articles) and a function that determines which source produces which items (or which items belong to which source) (e.g. which articles are published by which source). A size-frequency function f then measures, for every $n \in \mathbb{N}$, the number $f(n)$ being the number of sources with n items. In this paper we will adopt the continuous setting where, for $j > 0$, $f(j)$ denotes the density of sources with item density j (cf. Egghe (1990), Egghe and Rousseau (1990)). The most classical example of a size-frequency function f is a decreasing power law:

$$f(j) = \frac{C}{j^{\alpha}} \tag{1}$$

where C, $\alpha > 0$ are parameters. In informetrics, this regularity is called the law of Lotka, based on its introduction in Lotka (1926). Of course, a size-frequency function does not have to be of type (1). Any convexly decreasing function (such as an exponentially decreasing function) is a potential model for the size-frequency function f. A power function, however, is characterized as a scale-free function as follows (see Roberts (1979), Egghe (2004a,b)).

Definition I.1: a continuous function $f: \mathbb{R}^{+} \to \mathbb{R}^{+}$ is called scale-free if for every positive constant C there exists a positive constant D (only dependent on C) such that, for every $x \in \mathbb{R}^{+}$ we have

$$f(Cx) = Df(x) \tag{2}$$

The name scale-free is rightly choosen since a change of scale of the variable x (i.e. from x to Cx), yields the same function f, up to a scale factor D. The following important result is well-known - see Roberts (1979), Egghe (2004a) - where also a complete proof is given.

Theorem I.2: The following assertions are equivalent for a continuous function $f: ¡^{+} \to ¡^{+}$

(i) f is scale-free

(ii) there exist constants $a \in ¡^{+}, c \in ¡$ such that

$$f(x) = ax^{c} \tag{3}$$

for all $x \in ¡^{+}$.

It is clear that, when f is decreasing, we hence have a characterization of functions of the type (1). Since it is a characterization we also have that no other function (other than a power function) is scale-free. This is easily verified for an exponential function $f(x) = a^{x}$. Here

$$f(Cx) = a^{Cx} = \left(a^{C}\right)^{x} = b^{x}$$

which is, for $C \neq 1$, a function different from f (since $b = a^{C} \neq a$ and no constant D can be found such that $b^{x} = Da^{x}$ for all $x > 0$).

This unique property of Lotka's function (1) makes Lotkaian informetrics superior above other informetrics models. Indeed, as described in Egghe (2004a,b), Lotkaian informetric systems can be considered as self-similar fractals (for more on fractals we refer the reader e.g. to Feder (1988)) with fractal dimension $D = \alpha - 1$, where α is Lotka's exponent appearing in (1). So α is a direct measure of the fractal complexity of such IPPs.

Another important feature of Lotka's law is the fact that it is equivalent with other well-known laws as e.g. the law of Zipf (in linguistics), Pareto (in econometrics) and other ones (ee Egghe and Rousseau (1990), Egghe (2004a)). As a consequence of this, Lotka's law is found to be valid in many applications in these fields but also in all sorts of networks (e.g. with respect to the number of hyperlinks (in or out) or site sizes) such as intranets, the internet and WWW, citation networks and collaboration networks (see Egghe (2004a) or Bilke and Peterson (2001), Jeong, Tombor, Albert, Ottval and Barabási (2000), Barabási, Jeong, Néda,

Ravasz, Schubert and Vicsek (2002), Adamic, Lukose, Puniyani and Huberman (2001) and Barabási and Albert (1999)).

The scale-free property (characterizing power laws as described above) has also the following application e.g. in WWW concerning website sizes (i.e. the number of pages in a site): if we look at the distribution of site sizes for one arbitrary range, say sites that have between 1,000 and 2,000 pages, it would look the same as that for a different size range, say from 10 to 100 pages. In other words, "zooming" in or out in the scale at which one studies the web, one keeps obtaining the same result, just as in the case of zooming in or out on a self-similar fractal (cf. Huberman (2001)).

In the next section we will state and prove another charaterization of the scale-free property of size-frequency functions in terms of systematic sampling in sources and items of an IPP. Hence, in view of Theorem I.2, this will then also be a characterization of Lotka's law in terms of these sampling methods.

II. Systematic sampling in items and sources and their relation with scale-free information systems, i.e. Lotkaian informetrics.

Although we will define, in a mathematically correct way, systematic sampling in items and sources for continuous IPPs (and in terms of the function (1) for continuous variables $j > 0$), we will explain first the systematic sampling procedure "in practise", i.e. where we have a concrete IPP (e.g. a bibliography) with a finite (hence discrete) set of sources and items. Here we will reveal characteristic properties, in terms of the discrete size-frequency function $n \hat{I} ¥ ® f(n)$, of structural sampling in sources and items, valid except for discrete side-effects (rounding-off errors or failure because of small discrete values). In fact, these observations will show the necessity of the formulation of systematic sampling in sources and items in the continuous setting, where such drawbacks are not existing.

II.1 Systematic sampling in items and sources in practise

Systematic sampling (also called perfectly stratified sampling) in a finite set A of objects means the "taking" or "deletion" of every k^{th} object in the set ($k \hat{I} ¥$). In case of the "taking" we hence have a sample size which is a fraction $\frac{1}{k}$ of the size of the set A. In case of the "deletion" (hence taking the complement) we have a sample size which is a fraction $\frac{k-1}{k}$ of the size of the set A. Both values $\frac{1}{k}$ and $\frac{k-1}{k}$ are approximate but are exact if #A is a k-multiple (#A = the cardinality of A) and hence these values are exact, up to a rounding-off error which becomes neglectable for large values of #A. Note that, although the value k=1 is allowed this does not yield a real sample since the "taking" yields the original set A and the "deletion" yields the empty set Æ. If we talk about "every k^{th} object" this presupposes that the set A is ordered; if not, objects can be taken or deleted in a random way.

IPPs are more complex than a simple finite set: they consist of sources which have (or produce) items and hence we have to indicate, in case of sampling, where the sample is executed: in the items or in the sources. We will consider both. In the case of systematic sampling "taking or deleting every k^{th} object, i.e. item or source" we should also indicate the order in the set of items or sources. There are, for sampling in sources as well as for sampling in items, two clear options. For sources, we can use the order from most productive source to least productive source, i.e. start sampling from the sources which have the most items, i.e. following the order that is also used in the definition of the rank-frequency function, e.g. Zipf's law, see Egghe and Rousseau (1990) or Egghe (2004a). Alternatively one can use the opposite rank, i.e. starting with the least productive source. For systematic sampling in the items one can consider the same options, by using the order on the items, induced by the sources, or vice-versa.

In the sequel it will be clear that, apart from side-effects, both possibilities yield the same sample IPP in the sense of size-frequency function. In addition, these side-effects disappear when using continuous IPPs as we will do in the next subsection II.2.

II.1.1 Discussion on systematic sampling in items

When we make a systematic sample in the items, say with a sample fraction $\theta \in]0,1]$ (in the discrete setting, only rational values as indicated above are possible) it is clear that sources with a high number n of items will have a number of items, after the sample, close to θn (the higher n the correcter the value θn will be). Denoting by f the size-frequency function of the IPP before the sample and by f^* the size-frequency function after the sample, we hence have that, for large n,

$$f^*(\theta n) \approx f(n). \tag{4}$$

This relation (and the value θn) will become more exact the higher the value of n. Dependent on θ (e.g. $\theta = \frac{1}{2}$) relation (4) is also valid for small n (e.g. for $\theta = \frac{1}{2}$, (4) is also valid for $n = 2,4,...$ and there are rounding-off errors for the other (smaller) n-values.

This discussion makes it clear that relation (4) (with equality sign) is the fundamental relation for defining systematic sampling in the items in the continuous case.

II.1.2 Discussion on systematic sampling in sources

When we make a systematic sample in the sources, say with a sample fraction $\eta \in]0,1]$ (again, in the discrete setting, only rational values are possible) it is clear that sources (if used in the sample) keep their number n of items but that their number (i.e. $f(n)$) is reduced with a factor η (again apart from rounding-off errors which now might occur for high n since $f(n)$ is then low). So here we have that for large values of $f(n)$ (i.e. low n) we have (and $\approx$ is closer to = the higher $f(n)$)

$$f^*(n) \approx \eta f(n). \tag{5}$$

As above, this relation is also correct for low values of $f(n)$ (i.e. high n), for special values of η.

Note that the heuristic conclusions in Subsections II.1.1 and II.1.2 are independent on whether we use the "classical" source and item ranking (i.e. from most productive source on) or the reverse ranking. In Egghe (2002) we calculated concentration values of these sampled IPPs (e.g. systematic or truncated) and there the results were completely dependent on the used ranking: there only sampling in items or sources, in each case starting with the lower productive sources (or items in these sources) yields a higher concentration (inequality) structure than before the sampling while no results could be proved if the opposite ranking was used.

All the rounding-off effects discussed above in the discrete case do not play a role in the continuous case so that we now have an exact mathematical methodology to define systematic sampling in items and sources.

II.2 Systematic sampling in items and sources in continuous IPPs

Suppose we have given a continuous IPP, i.e. an IPP in which we have a size-frequency function $f(j)$, dependent on the continuous variable $j \in \mathbb{R}^+$. Let us denote by f^* the size-frequency function of a sampled IPP.

Definition II.2.1

A sample is a systematic sample in the items (or an item systematic sample), with sample fraction θ (i.e. the sample size is θA, a fraction θ of all the items) if, for every $j \in \mathbb{R}^+$:

$$f^*(\theta j) = f(j). \tag{6}$$

Definition II.2.2

A sample is a systematic sample in the sources (or a source systematic sample), with sample faction η (i.e. the sample size is ηT, a fraction η of all the sources) if, for every $j \in \mathbb{R}^+$:

$$f^*(j) = \eta f(j). \tag{7}$$

With these exact definitions we are able to state and prove the following result, characterizing Lotkaian informetrics.

Theorem II.2.3: The following assertions (i) and (ii) are equivalent:

(i) Item and source systematic samples are the same, i.e.:

(a) For every $\theta \in]0,1]$ there exists a $\eta \in]0,1]$ (only dependent (injectively) on θ) such that every item systematic sample with fraction θ is a source systematic sample with fraction η. Reversely:

(b) For every $\eta \in]0,1]$ there exists a $\theta \in]0,1]$ (only dependent (injectively) on η) such that every source systematic sample with fraction η is an item systematic sample with fraction θ.

(ii) The function f is scale-free and hence, equivalently (Section I), f is a decreasing power law (i.e. Lotka's law (1)).

If the above assertions are true, the relation between η and θ is given by

$$\eta = \theta^{\alpha} \tag{8}$$

where α is Lotka's exponent, see (1).

Proof:

(i) $\Rightarrow$ (ii)

Suppose (i)(a). We have given formula (6) and hence, by (i)(a), we also have formula (7) with $\eta = \eta(\theta)$, i.e. η is an injective function of θ. So, for all $j \in \mathbb{N}^{+}$:

$$f^{*}(\theta j) = f(j)$$

and

$$f^{*}(\theta j) = \eta f(\theta j).$$

Hence

$$f(\theta j) = \frac{1}{\eta} f(j) \tag{9}$$

for all $j \in \mathbb{R}^{+}$, where $\eta = \eta(\theta)$ for all $\theta \in]0,1]$.

Suppose (i)(b). We have given formula (7) and hence, by (i)(b), we also have formula (6) with $\theta = \theta(\eta)$, i.e. θ is an injective function of η. So, for all $j \in \mathbb{R}^{+}$:

$$f^{*}(j) = \eta f(j)$$

$$f^{*}(j) = f^{*}\left(\theta \frac{1}{\theta} j\right) = f\left(\frac{1}{\theta} j\right).$$

Hence

$$f\left(\frac{1}{\theta} j\right) = \eta f(j) \tag{10}$$

for all $j \in \mathbb{R}^{+}$.

Note that the fact that $\theta = \theta(\eta)$ is an injection implies that η is a function of $\frac{1}{\theta}$, for all $\theta \in]0,1]$. Hence, by (9) and (10), f is scale-free (since $\theta \in]0,1]$ and $\frac{1}{\theta} \in [1,+\infty[$ and hence all values in $\mathbb{R}^{+}$ are covered by Definition I.1) and the result follows.

(ii) $\Rightarrow$ (i)

Let f be given as a decreasing power law

$$f(j) = \frac{C}{j^{\alpha}}$$

$j \in \mathbb{N}^+$. Define, given $\theta \in (0,1]$

$$f^*(\theta j) = f(j)$$

(i.e. given an item systematic sample) for all $j \in \mathbb{N}^+$. Hence, for all $j \in \mathbb{N}^+$:

$$f^*(j) = f\left(\frac{1}{\theta} j\right)$$

$$f^*(j) = \theta^{\alpha} f(j)$$

Putting $\eta = \theta^{\alpha}$ we see that we have a source systematic sample (with fraction $\eta = \theta^{\alpha}$ (note that indeed $\theta \in (0,1]$ implies $\eta \in (0,1]$ and that η is an injective function of θ)). Hence we proved (i)(a).

Let us now have a given source systematic sample (i.e. given $\eta \in (0,1]$)

$$f^*(j) = \eta f(j)$$

for all $j \in \mathbb{N}^+$. Hence, for all $j \in \mathbb{N}^+$:

$$f^*\left(\eta^{\frac{1}{\alpha}} j\right) = \eta f\left(\eta^{\frac{1}{\alpha}} j\right)$$

$$= \eta \frac{f(j)}{\eta}$$

$$f^*\left(\eta^{\frac{1}{\alpha}} j\right) = f(j)$$

Hence we have an item systematic sample with fraction $\theta = \eta^{\frac{1}{\alpha}}$ (note again that $\eta \in]0,1]$ implies $\theta \in]0,1]$ and that θ is an injective function of η). Hence we proved (i)(b). This completes the proof of the theorem. ~

The above result shows that Lotkaian informetrics (and only Lotkaian informetrics) allows for a sample size-frequency function f^* which is (up to constants) the same as the population size-frequency function f, a remarkable conclusion.

References

L.A. Adamic, R.M. Lukose, A.R. Puniyani and B.A. Huberman (2001). Search in power-law networks. Physical Review E64, 46135-46143, 2001.

A.-L. Barabási, and R. Albert (1999). Emergency of scaling in random networks. Science 286, 509-512, 1999.

A.-L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert and T. Vicsek (2002). Evolution of the social network of scientific collaborations. Physica A311, 590-614, 2002.

S. Bilke and C. Peterson (2001). Topological properties of citation and metabolic networks. Physical Review E6403(3), 76-80, 2001.

L. Egghe (1990). The duality of information systems with applications to the empirical laws. Journal of Information Science 16(1), 17-27, 1990.

L. Egghe (2002). Sampling and concentration values of incomplete bibliographies. Journal of the American Society for Information Science and Technology 53(4), 271-281, 2002.

L. Egghe (2004a). Lotkaian Informetrics. Book in preparation.

L. Egghe (2004b). The power of power laws and the interpretation of Lotkaian informetric systems as self-similar fractals. Journal of the American Society for Information Science and Technology, to appear, 2004.

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

J. Feder (1988). Fractals. Plenum, New York, USA, 1988.

B.A. Huberman (2001). The Laws of the Web. Patterns in the Ecology of Information. The MIT Press, Cambridge (MA), USA, 2001.

H. Jeong, B. Tombor, R. Albert, Z.N. Ottval and A.-L. Barabási (2000). The large-scale organization of metabolic networks. Nature 407, 651-654, 2000.

A.J. Lotka (1926). The frequency distribution of scientific productivity. Journal of the Washington Academy of Sciences 16(12), 317-324, 1926.

F.S. Roberts (1979). Measurement Theory with Applications to Decisionmaking, Utility, and the social Sciences. Addison-Wesley, Reading (MA), USA, 1979.