Sampling and concentration values of incomplete bibliographies

Non Peer-reviewed author version

EGGHE, Leo (2002) Sampling and concentration values of incomplete bibliographies. In: Journal of the American Society for Information Science and Technology, 53(4). p. 271-281.

DOI: 10.1002/asi.10033
Handle: http://hdl.handle.net/1942/775

Sampling and concentration values of incomplete bibliographies

by

L. Egghe, LUC, Universitaire Campus, B-3590 Diepenbeek, Belgium[1]
and
UIA, Universiteitsplein 1, B-2610 Wilrijk, Belgium
e-mail : leo.egghe@luc.ac.be

ABSTRACT

This paper studies concentration aspects of bibliographies. More in particular we study the impact of incompleteness of such a bibliography on its concentration values (i.e. its degree of inequality of production of its sources). Incompleteness is modelled by sampling in the complete bibliography. The model is general enough to comprise truncation of a bibliography as well as a perfectly stratified sample on sources or items. In all cases we prove that the sampled bibliography (or incomplete one) has a higher concentration value than the complete one. These models hence shed some light on the measurement of production inequality in incomplete bibliographies.

[1]Permanent address

Acknowledgement : the author is grateful to prof. Dr. R. Rousseau for interesting discussions on the topic of this paper.

I. Introduction

Bibliographies (or its generalization : Informetric Production Processes (IPPs)) are formally described e.g. in Egghe (1989,1990). A bibliography consists of a set of sources (e.g. journals), a set of items (e.g. articles) and a function that points out which items belong to which source (i.e. pointing out, for each article, in which journal it is published). Other interpretations of IPPs (even beyond the information sciences, e.g. in econometrics, biometrics, etc.) exist (see e.g. Egghe and Rousseau (1990)) but we will not use them here. We will henceforth use the terminology : bibliography.

Typical for all bibliographies is the large inequality that exists between the production of the sources. Intuitively speaking, few sources have many items and many sources have few items. One talks in this connections also about "elitism" or "elitarism" of bibliographies - also found e.g. in econometrics models relating to richness and poverty. In informetrics, the formal way to express inequality is given by the classical informetric laws such as the ones of Lotka, Zipf, Mandelbrot, Bradford, Leimkuhler and so on (see e.g. Egghe and Rousseau (1990)). The measurement of inequality can be performed using concentration theory, see Egghe and Rousseau (1990), Rousseau (1992) but its invention goes back to the beginning of the 20th century, see e.g. Lorenz (1905), Gini (1909). In its most elementary way it goes as follows. We start from a vector

$$X = (x_1,\ldots, x_N)$$

where, for each $i=1,\ldots,N$, $x_i \geq 0$ denotes the number of items in (more generally the production of) the i^{th} source in the bibliography. This sequence is ordered in a monotone way - here we will order it decreasingly. We transform X into its corresponding vector of relative values :

$$A_X = (a_1,\ldots,a_N)$$

where

$$a_i = \frac{x_i}{\sum_{j=1}^{N} x_j} \qquad (1)$$

for every i=1,...,N. The Lorenz curve L_X of X is then formed by linearly interconnecting the points in the unit square

$$(0,0), (\frac{1}{N},a_1), (\frac{2}{N},a_1+a_2), ..., (\frac{i}{N},\sum_{j=1}^{i} a_j), ..., (1,1)$$

Note that indeed $\sum_{j=1}^{N} a_j=1$. Since X is decreasing, the Lorenz curve L_X is concavely increasing between (0,0) and (1,1). Let

$$Y = (y_1,...,y_N)$$

be a second vector as above. We say that Y represents a more concentrated situation than X if

$$L_Y \geq L_X \qquad (2)$$

and where the higher concentration is strict if $L_Y>L_X$ at least in some points of the graphs.

One can then proceed with this classical concentration theory by defining good concentration measures M as functions on such vectors X, Y such that $L_Y>L_X$ implies M(Y)>M(X). Classical examples of good concentration measures are : V (the coefficient of variation), G (the Gini index), P (Pratt's measure), Th (Theil's measure), see Egghe and Rousseau (1990), but we will not be using these measures here. It is well-known that comparing Lorenz curves is the perfect solution to comparing inequality.

The above method applies to any bibliography : here in $X = (x_1,...,x_N)$, each x_i represents the number of articles in the i^{th} journal in the bibliography (where journals are ordered in decreasing order of the number of articles they have (on a certain subject)). In practise, however, we are faced with the problem of incomplete bibliographies : we never know the complete bibliography, e.g. due to the imperfectness of information retrieval machines. This incompleteness can also be interpreted in the following natural case : suppose we want to keep track of a bibliography in time. Then the cumulative set of journals and articles up to, say 1999, is an incomplete version (or extract or sample) of the one up to, say 2000. Interpreted in

this way, incomplete bibliographies have their application in the study of (even complete) bibliographies in function of time. Hence, the term "incomplete" does not only have to be interpreted in its "smallest" interpretation, i.e. the one in which we only retrieve a part of what we want : the complete bibliography.

An incomplete version of a bibliography can be interpreted in several ways. In any way we will consider the incomplete version of a bibliography as a bibliography that is obtained as a sample in the original one. Of course, there are many ways to execute a sample. Because of the dual character of bibliographies, at least two "major" types of samples are possible : a sample in the items and a sample in the sources. We refer to Rousseau (1993) for a first attempt to study the effect of sampling on the concentration properties of a bibliography, based on elaborated (theoretical) examples. In addition to these two types we can - in case we end up with a source with zero items (in case of an items sample) or with a non-picked source (in case of a source sample) :

1. allow this source but with zero items (x_i=0), or
2. delete this source, i.e. considering as non-existent.

These two approaches are considerably different. Without going into the different sampling techniques discussed in the sequel, let us illustrate this by a simple example. Suppose

$$X = (5,4,3,2,1)$$

is our "complete" situaton, hence a bibliography where we have one source with 5,4,3,2,1 items, respectively. Deleting the items in the sources with 1 and 2 items leaves us two possibilities, as described above

1. The sampled bibliography is represented by the vector, denoted by s(X) :
$$s(X) = (5,4,3,0,0)$$
2. The sampled bibliography is represented by the vector, denoted by $\sigma(X)$:
$$\sigma(X) = (5,4,3)$$

These two cases are, conceptually, very different. Denoted in a general way, the difference between

$$(x_1,\ldots,x_{N-k},\underbrace{0,\ldots,0}_{k})$$

and

$$(x_1,\ldots,x_{N-k})$$

$(x_1,\ldots,x_{N-k}>0)$ can be the difference between (econometric interpretation)

1. a group of N persons where the first N-k ones have a good salary, while the last k persons do not earn any money

and

2. a group of N-k persons, all having a good salary.

In an informetric interpretation, we have the above difference :

1. we have a group of N researchers (in a scientific domain) in which N-k of them are very productive (in terms of number of publications) and in which k of them are not-productive at all

and

2. we have a group of N-k very productive researchers.

The above arguments lead to the following 4 sampling types (many more methods will be explained in the sequel - we do not go into this now)

1. sampling in items, keeping zero sources (or in other terms : the number of sources (N) is fixed),
2. sampling in items, deleting zero sources (here the number of sources varies),

3. sampling in sources, keeping not-selected sources as zero sources (again using a fixed number N of sources),
4. sampling in sources, deleting not-selected sources (again here the number of sources varies).

The ultimate goal of studying the above sampling types - besides their proper theoretical interest - is to be able to conjecture some results concerning the concentration of a real incomplete bibliography, e.g. a retrieved bibliography and to determine how it differs from the concentration of a (unknown) bibliography.

The next section deals with sampling in items. There we prove, using a very general sampling method (to be discussed there and comprising truncation of a bibliography as well as perfectly stratified samples of the bibliography - see further for exact definitions) that, if the number of sources is fixed and if we sample from the least productive sources to the most productive ones (the most important case as will be explained there), the Lorenz curve of the sampled bibliography is always above the one of the complete bibliography. In all other cases (including the deletion of zero-sources) we produce counterexamples showing that the Lorenz curve of the sampled bibliography is not always above or below the one of the complete bibliography.

The third section deals with sampling in sources. Also here we prove that, if the number of sources is fixed and if we keep the most productive sources (see further for an exact definition), the Lorenz curve of the sampled bibliography is always above the one of the complete bibliography.

In summary, the most "natural" sampling types lead to an increase of the inequality (in production of the sources) in the sampled bibliography. This hence leads to a systematic over-estimation of the inequality (concentration) of the complete bibliography.

II. Sampling items

We will first introduce two important item sampling methods which will turn out to be two extreme cases of the general sampling method that we will discuss in this section.

II.1. Perfectly stratified sample (PSS)

As always, a bibliography is represented by a production vector

$$X = (x_1,\ldots,x_N)$$

where $x_i \in \mathbb{N}\cup\{0\}$, for all i=1,...,N. We order X decreasingly and sample in the items, using the least productive sources first. Let $\theta \in Q^+$ (the positive rational numbers) be such that $0 \le \theta \le 1$ and that $\sum_{j=1}^{N} x_j$, the total number of items, is a $\frac{1}{\theta}$ - multiple (this is used in order to have no rounding - off errors which would disturb the results as we will see in the sequel). A perfectly stratified sample (PSS) starts replacing x_N by $[\theta x_N]$ ([x] denotes the largest entire number, smaller than or equal to x). This will be the N^{th} coordinate $s(x_N)$ of the sampled vector, denoted by s(X). The "rest" $\frac{1}{\theta}(\theta x_N - [\theta x_N])$ is then added to x_{N-1} and, for the $(N-1)^{th}$ coordinate of s(X) we take

$$s(x_{N-1}) = [\theta(x_{N-1} + \frac{1}{\theta}(\theta x_N - [\theta x_N]))] \qquad (3)$$

The rest is added to x_{N-2} and so on. The notation is getting rather complicated but there is a simple way to express the cumulative number $\sum_{j=i}^{N} s(x_j)$ for every i=1,...,N. It is nothing else than

$$\sum_{j=i}^{N} s(x_j) = [\theta(\sum_{j=i}^{N} x_j)] \qquad (4)$$

which makes life much more easy and allows us not to use (3) further on : (4) will suffice.

Another way to explain PSS is as follows. Let $\theta = \frac{1}{n}$ ($n \in \mathbb{N}$). Replace each x_j by a unit row vector of length x_j , so that the production vector $(x_1,\ldots,x_N)$ is replaced by $((1,\ldots,1)(1,\ldots,1)(\ldots)(1,\ldots,1))$. Ignoring the inner brackets we essentially have a unit vector of length $\sum_{j=1}^{N} x_j$. *Choosing our* $\frac{1}{n}$*-sample is then just a matter of going along this vector* (from right to left), ignoring the inner brackets and highlight each n^{th} entry. Then $s(x_j)$ is just the number of highlighted entries in the bracket corresponding to the j^{th} source.

The same sort of construction can be described when $\theta = \frac{p}{q}$ ($p,q \in \mathbb{N}$, $p<q$). Here the production vector $(x_1,\ldots,x_N)$ is replaced by replacing each x_j by a unit matrix with p rows and x_j columns so that overall we have a unit matrix with p rows and $\sum_{j=1}^{N} x_j$ columns. To execute the $\frac{p}{q}$-sample we run down the columns successively highlighting ever q^{th} entry and then count up the number of highlighted entries within each matrix.

Some examples will show the simplicity of the method.

1. $X = (4,1,1) = (x_1,x_2,x_3)$ and $\theta = \frac{1}{2}$. Since $\frac{1}{2} < x_3 < 1$ we take $s(x_3)=0$ and we shift $x_3=1$ to the second and add it. Now $2 \cdot \frac{1}{2} = 1$ yielding $s(x_2) = 1$. Finally $4 \cdot \frac{1}{2} = 2$ yielding $s(x_1) = 2$. Hence $s(X) = (2,1,0)$. Note that $\sum_{j=1}^{3} x_j = 6$ is a $\frac{1}{\theta} = 2$ - multiple.

   In the description with the unit vectors we rewrite X as

   $$((1,1,1,1),(1),(1))$$

   and highlight every second 1 from the right. So we have

   $$((\underline{1},1,\underline{1},1),(\underline{1}),(1))$$

   yielding $s(X)=(2,1,0)$ again.

   In the same way the following exercices can be executed.

2. $X = (4,2,1,1,1)$, $\theta = \frac{1}{3}$. Note that $\sum_{j=1}^{5} x_j = 9 = 3$ - multiple. Applying the same method now yields (verify)

   $$s(X) = (2,0,1,0,0)$$

but since s(X) is not decreasing we define (order s(X) decreasingly)

$$s'(X) = (2,1,0,0,0)$$

θ can be any number in $Q^+ \cap [0,1]$ such that $\sum_{i=1}^{N} x_i$ is a $\frac{1}{\theta}$ - multiple and not just a number of the form $\theta = \frac{1}{n}$, $n \in \mathbb{N}$. Example :

3. $X = (1,1,1,1,1)$, $\theta = \frac{3}{5}$. Note that $\sum_{j=1}^{5} x_j = 5$, a $\frac{5}{3}$ - multiple. Verify that

$$s(X) = (1,1,0,1,0)$$

and hence that

$$s'(X) = (1,1,1,0,0).$$

PSS is one of the most important sampling methods, which is also applied in real-life to have a fast method that resembles (or approximates) random sampling (see e.g. Clarke and Cooke (1992), Carpenter and Storey Vasu (1978) or Egghe and Rousseau (1990, 2001a)). Only in cases of production units in factories, where a production error might occur in every n^{th} object say, PSS might give sampling results which differ from random sampling. As described in the mentioned references, there is very little chance that we have this problem in sampling in bibliographies. In our interpretation we think it resembles the "making" of incomplete bibliographies very well. Yet, PSS will be generalized in this section, where source-variable θ will be allowed (see further).

II.2. Truncation

Let the bibliography be represented by $X=(x_1,...,x_N)$, ordered decreasingly. Let $i \in \{1,...,N\}$. The i-truncation of this bibliography is obtained by keeping the i most productive sources and putting the sources on rank i+1,...,N on zero production. Hence the i-truncation of the bibliography is represented by

$$s(X) = (x_1,\ldots,x_i,\underbrace{0,\ldots,0}_{N-i}) \quad (5)$$

The i-truncation can be considered as the bibliography consisting of the i "core" sources of the original bibliography (see Egghe and Rousseau (2001b) for a treatment of cores of a bibliography).

II.3. General model for sampling in items.

The above item sampling methods can be generalized as follows. Let $X=(x_1,\ldots,x_N)$ represent our bibliography. We suppose X to be decreasing. The philosophy of this general method is allowing for variable sample fractions θ_i (i=1,...,N) dependent on the source i. Most naturally we require $(\theta_i)_{i=1,\ldots,N}$ to be decreasing (including a constant sequence) : in this sampling method, items in low productive sources (high i) have a lower chance to be picked for the sample than items have in sources with low i (highly productive sources). In exact mathematical terms, $s(X)=(s(x_1),\ldots,s(x_N))$, the representation of the sampled bibliography, is obtained as follows :

$$s(x_N) = [\theta_N x_N] \quad (6)$$

, of course sampling first in the low productive sources. The decimal rest, $\theta_N x_N - [\theta_N x_N]$ is then added to the $(N-1)^{th}$ source (the same as in II.1). Then

$$s(x_{N-1}) = [\theta_{N-1}x_{N-1} + \theta_N x_N - [\theta_N x_N]] \quad (7)$$

and so on. The simplest way to describe this sampling model is as follows : for every i=1,...,N

$$\sum_{j=i}^{N} s(x_j) = [\sum_{j=i}^{N} \theta_j x_j] \quad (8)$$

From this it also follows that, for every i=1,...,N

$$s(x_i) = [\sum_{j=i}^{N} \theta_j x_j - \sum_{j=i+1}^{N} s(x_j)] \quad (9)$$

since $\sum_{j=i+1}^{N} s(x_j) \in \mathbb{N}$. Formula (9) is easily obtained and certainly easier than deriving it from a complete induction argument based on (6) and (7) (which, however, is also possible). The vector representing the sampled bibliography is denoted $s(X)=(s(x_1),\ldots,s(x_N))$. Its decreasing version is denoted $s'(X)=(s'(x_1),\ldots,s'(x_N))$. Denote by L_X, $L_{s(X)}$, $L_{s'(X)}$, the Lorenz-curves derived from X, s(X), s'(X). We have the following general result.

Theorem II.3.1. If $(\theta_i)_{i=1,\ldots,N}$ is decreasing and if $\sum_{j=1}^{N} \theta_j x_j \in \mathbb{N}$, then

$$L_{s'(X)} \geq L_{s(X)} \geq L_X \tag{10}$$

Hence the sampled bibliography is more concentrated than the original one.

Since the proof is a bit lengthy we give it in Appendix 1.

Note that the sampling methods described in II.1 and II.2 are a special case of this

1. PSS is obtained by taking

$$\theta_1 = \theta_2 = \ldots = \theta_N = \theta \tag{11}$$

Note that in this case, the condition that $\sum_{j=1}^{N} x_j$ must be an $\frac{1}{\theta}$ - multiple is the same as the requirement in Theorem II.3.1 above :

$$\sum_{j=1}^{N} \theta_j x_j \in \mathbb{N} \tag{12}$$

This is a very natural requirement where we only look at bibliographies in which "entire" items are sampled. Besides, below, we will give a counterexample to Theorem II.3.1 in case (12) is not satisfied. So, since PSS is included in Theorem II.3.1 we have here also that the sampled bibliography is more concentrated than the original one.

2. Truncation is obtained by taking

$$\begin{cases} \theta_1 = \ldots = \theta_i = 1 \\ \theta_{i+1} = \ldots = \theta_N = 0 \end{cases} \tag{13}$$

Note that here (12) is always valid and that s'(X)=s(X). Since Theorem II.3.1 applies to truncation, we have a generalization of the result obtained in Egghe and Rousseau (2001b) which has, as mentioned above, an application in the determination of the core of a bibliography.

II.4. Example showing that the requirement (12) cannot be dropped in Theorem II.3.1.

Even in the case of a PSS we can give a counterexample. Take X=(3,1,1), θ=0.5, hence $\sum_{j=1}^{3} x_j = 5$ is not a $\frac{1}{\theta} = 2$ - multiple. It is easily seen that s(X)=(1,1,0), hence

$$A_X = \left(\frac{3}{5}, \frac{1}{5}, \frac{1}{5}\right)$$

$$A_{s(X)} = \left(\frac{1}{2}, \frac{1}{2}, 0\right)$$

showing that L_X and $L_{s(X)}$ cross.

If we sample items, using the most productive sources first, Theorem II.3.1 is not true.

II.5. If the sampling method II.3 is done in the reverse way, i.e. by starting with the most productive sources, then Theorem II.3.1 is not true, nor do we have an opposite result.

This sampling method replaces (6)-(8) by

$$s(x_1) = [\theta_1 x_1]$$

$$s(x_2) = [\theta_2 x_2 + \theta_1 x_1 - [\theta_1 x_1]]$$

$$\sum_{j=1}^{i} s(x_j) = \left[\sum_{j=1}^{i} \theta_j x_j\right]$$

Even in the PSS-case ($\theta_1 = \ldots = \theta_N = \theta$) the analogue of Theorem II.3.1, nor the opposite result ($L_{s(X)} \leq L_X$) are generally true. Indeed, take X=(5,1,1,1,1), $\theta = \frac{1}{3}$. Then this sampling method yields s(X)=(1,1,0,0,1) and hence s'(X)=(1,1,1,0,0). But

$$A_X = \left(\frac{5}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}\right),$$

$$A_{s'(X)} = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0, 0\right)$$

showing that $L_{s'(X)}$ and L_X cross and hence also $L_{s(X)} \not\geq L_X$.

II.6 If the sampling method II.3 is applied but with dropping the zero sources, then Theorem II.3.1 is not true, nor do we have an opposite result.

Take X=(12,4,2,1,1), θ=0.25. Then here s(X)=(3,1,1) and so

$$A_X = \left(\frac{12}{20}, \frac{4}{20}, \frac{2}{20}, \frac{1}{20}, \frac{1}{20}\right),$$

$$A_{s(X)} = \left(\frac{3}{5}, \frac{1}{5}, \frac{1}{5}\right)$$

It is trivial to see that $L_{s(X)} < L_X$. Of course, any example where no zero sources occur yields $L_{s(X)} \geq L_X$, by Theorem II.3.1, since this sampling method and the one of II.3 then coincide. The next is an example where $L_{s(X)}$ and L_X cross.

Take X=(3,3,2,2,1,1), θ=0.5. Then s(X)=(2,1,1,1,1),

$$A_X = \left(\frac{3}{12}, \frac{3}{12}, \frac{2}{12}, \frac{2}{12}, \frac{1}{12}, \frac{1}{12}\right),$$

$$A_{s(X)} = \left(\frac{2}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6}\right).$$

The equation of the line connecting (0,0) with $\left(\frac{2}{6},\frac{6}{12}\right)$ is $y=\frac{3}{2}x$. For $x=\frac{1}{5}$ this yields $L_X\left(\frac{1}{5}\right)=\frac{3}{10}$ while $L_{s(X)}\left(\frac{1}{5}\right)=\frac{2}{6}>\frac{3}{10}$. But $L_X\left(\frac{2}{5}\right)>L_X\left(\frac{2}{6}\right)=\frac{1}{2}=L_{s(X)}\left(\frac{2}{5}\right)$. This shows that L_X and $L_{s(X)}$ cross.

This completes the study of all cases of sampling on items. We now proceed with the case of sampling on sources.

III. Sampling sources.

Also here, the most important case of sampling in sources, namely perfectly stratified sampling, will yield a result as in Theorem II.3.1, namely $L_{s'(X)}\geq L_X$. It is the case where we keep the non-picked sources as zero-sources and where we sample in such order that the largest sources are kept in the sample. This will be described now.

III.1 Description of the model of sampling in sources.

Again we represent the bibliography by $X=(x_1,\ldots,x_N)$ (decreasing) and we let $\eta\in Q^+\cap]0,1[$ such that N is an $\frac{1}{\eta}$ - multiple. A perfectly stratified sample (PSS) in sources, giving priority to higher productive sources yields the bibliography represented by the vector

$$s(X) = \left(x_1,\ldots,x_{\frac{1}{\eta}-1},0,x_{\frac{1}{\eta}+1},\ldots,x_{\frac{2}{\eta}-1},0,\ldots,x_{N-1},0\right) \tag{14}$$

The decreasing order version of this vector is then

$$s'(X) = \left(x_1,\ldots,x_{\frac{1}{\eta}-1},x_{\frac{1}{\eta}+1},\ldots,x_{\frac{2}{\eta}-1},\ldots,\ldots,x_{N-1},\underbrace{0,\ldots,0}_{K}\right) \tag{15}$$

where $N=\frac{K}{\eta}$, $K\in\mathbb{N}$.

We have the following result.

Theorem III.1.1. The above sampling method in sources yields

$$L_{s'(X)} \geq L_X. \tag{17}$$

Hence the sampled bibliograhy is more concentrated than the original one.

Since the proof is rather technical we give it in Appendix 2.

III.2 The above theorem is not valid for the (non-decreasing) s(X).

This is clear, since so many zeroes occur. An example

Take X=(4,3,2,1), η=0.5. Then s(X)=(4,0,2,0). Hence

$$A_X = \left(\frac{4}{10}, \frac{3}{10}, \frac{2}{10}, \frac{1}{10}\right)$$

and

$$A_{s(X)} = \left(\frac{4}{6}, 0, \frac{2}{6}, 0\right)$$

So $L_X\left(\frac{1}{4}\right) = \frac{4}{10} < L_{s(X)}\left(\frac{1}{4}\right) = \frac{4}{6}$ but $L_X\left(\frac{1}{2}\right) = \frac{7}{10} > L_{s(X)}\left(\frac{1}{2}\right) = \frac{4}{6}$.

III.3 The above theorem is not valid for not perfectly stratified samples.

Indeed, even the simplest case of replacing one source by a zero-source, does not yield the result. Take X=(3,3,1,1,1,1) and s'(X)=(3,1,1,1,1,0) (hence replacing one source with 3 items by a zero-source). Now

$$A_X = \left(\frac{3}{10}, \frac{3}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}\right)$$

and

$$A_{s'(X)} = \left(\frac{3}{7},\frac{1}{7},\frac{1}{7},\frac{1}{7},\frac{1}{7},0\right)$$

Note that $L_X\left(\frac{1}{6}\right)=\frac{3}{10}<L_{s'(X)}\left(\frac{1}{6}\right)=\frac{3}{7}$ but $L_X\left(\frac{2}{6}\right)=\frac{6}{10}>L_{s'(X)}\left(\frac{2}{6}\right)=\frac{4}{7}$. Note that L_X and $L_{s'(X)}$ intersect twice since $L_X\left(\frac{5}{6}\right)<L_{s'(X)}\left(\frac{5}{6}\right)=1$.

III.4 The above theorem is not valid if we sample in the reverse way.

Here, instead of replacing each $x_{\frac{i}{\eta}}$ (i=1,...,K) by 0 as in (14) we replace each $x_{N-\frac{i}{\eta}+1}$ by 0.

Example : X=(10,1,1,1), η=0.5, s'(X)=(1,1,0,0). But

$$A_X = \left(\frac{10}{13},\frac{1}{13},\frac{1}{13},\frac{1}{13}\right)$$

$$A_{s'(X)} = \left(\frac{1}{2},\frac{1}{2},0,0\right)$$

Hence L_X and $L_{s'(X)}$ cross.

III.5 The above theorem is not valid if we delete sources (instead of replacing them by 0).

Indeed, take X=(4,3,2,1), η=0.5 (and sample as in the theorem, except that we delete non-picked sources). Here s(X)=(4,2) and hence

$$A_X = \left(\frac{4}{10},\frac{3}{10},\frac{2}{10},\frac{1}{10}\right)$$

$$A_{s(X)} = \left(\frac{4}{6},\frac{2}{6}\right)$$

Hence $L_X\left(\frac{1}{2}\right)=\frac{7}{10}>L_{s(X)}\left(\frac{1}{2}\right)=\frac{4}{6}$.

Also an opposite example exists : X=(10,1,1,1), η=0.5. Then s(X)=(10,1), hence

$$A_X = \left(\frac{10}{13},\frac{1}{13},\frac{1}{13},\frac{1}{13}\right)$$

and

$$A_{s(X)} = \left(\frac{10}{11},1\right)$$

So $L_X<L_{s(X)}$ here.

IV. Conclusions.

In this paper we studied different sampling techniques in bibliographies, comprising sampling in items and sampling in sources.

When sampling in items we allow the probability for an item to be picked to be increasing with the number of items in the sources. In this general setting we prove that, if we start sampling in the least productive sources (hereby keeping zero-sources), the Lorenz curve of the sampled bibliography is above the Lorenz-curve of the original one. In other words, the sampled bibliography is more concentrated than the original one. The model and result applies to the case of perfectly stratified sampling (often used as an approximation for random sampling) as well as to truncation of bibliographies (i.e. only using the "core" of the bibliography - see Egghe and Rousseau (2001b)).

When sampling in sources, a similar result is proved. Here we show that a perfectly stratified sample in sources, keeping the most productive sources and replacing non-picked sources by a zero-source, yields a sampled bibliography for which the Lorenz-curve is above the one of the original bibliography.

We also show that all variants of the above methods (reversing the order, deleting zero-sources,...) do not yield such (or another) result.

So, in the two major sampling methods, we have that the sampled bibliography is more concentrated than the original one. This gives information about the concentration of bibliographies (as we receive them, e.g. as the result of an IR action) as compared to the (unknown) complete one. In all cases we can say that the observed concentration is higher than the concentration of the complete bibliography.

Note

As remarked by one of the referees, the requirement that $\sum_{j=1}^{N} \theta_j x_j \in \mathbb{N}$ (in case of sampling in items - Theorem II.3.1) and the requirement that N is an $\frac{1}{\eta}$-multiple (in case of sampling in sources - see subsection III.1) is not always valid in practise. Very simply, if N is prime, the sampling in sources is not even possible.

What is behind these requirements is that

(i) for exact results, we need them (since we show by example that without them the results are wrong)

(ii) for large N (which is always the case) one can drop the requirements, hereby only making a mistake in the last item (or source) sampled. We estimate that in this case (large N) the found increase in concentration will be there. In short, we indicate that, in practical cases, concentration increases.

Problem

We leave it as an open problem (for further study) to make explicite calculations of the difference of concentration between a bibliography and its sampled version. It would yield information on the concentration values of a complete (unknown) bibliography.

References

R.L. Carpenter and E. Storey Vasu (1978). Statistical Methods for Librarians. American Library Association, Chicago, USA.

G.M. Clarke and D. Cooke (1992). A basic Course in Statistics. Third Edition. Edward Arnold, London, UK.

L. Egghe (1989). The duality of informetric Systems with Applications to the empirical Laws. Ph. D. Thesis. The City University, London, UK.

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

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

L. Egghe and R. Rousseau (2001a). Elementary Statistics for effective Library and Information Service Management, Aslib, to appear.

L. Egghe and R. Rousseau (2001b). The core of a scientific subject : an exact definition using concentration theory and fuzzy set theory, to appear.

C. Gini (1909). Il diverso accrescimento delle classi sociali e la concentrazione della richezza. Giornale degli Economisti, serie 11, 37.

M.O. Lorenz (1905). Methods of measuring concentration of wealth. Journal of the American Statistical Association, 9, 209-219.

R. Rousseau (1992). Concentration and diversity Measures in informetric Research. Doctorate Dissertation, University of Antwerp.

R. Rousseau (1993). Measuring concentration : sampling design issues, as illustrated by the case of perfectly stratified samples. Scientometrics, 28(1), 3-14.

Appendix 1

Proof of Theorem II.3.1

That $L_{s'(X)} \geq L_{s(X)}$ is trivial since $s'(X)$ is the decreasing re-order of $s(X)$ and hence, for every $i=1,\ldots,N$

$$\sum_{j=1}^{i} s'(x_j) \geq \sum_{j=1}^{i} s(x_j) \tag{A1}$$

and note that $\sum_{j=1}^{N} s'(x_j) = \sum_{j=1}^{N} s(x_j)$.

Hence it suffices to prove that $L_{s(X)} \geq L_X$. Denote by T the total number of items, $\sum_{j=1}^{N} x_j$. L_X is obtained by linearly connecting the points

$$(0,0), \left(\frac{1}{N}, \frac{x_1}{T}\right), \ldots, \left(\frac{i}{N}, \frac{\sum_{j=1}^{i} x_j}{T}\right), \ldots, (1,1) \tag{A2}$$

$L_{s(X)}$ is obtained by linearly connecting the points

$$(0,0), \left(\frac{1}{N}, \frac{s(x_1)}{\sum_{j=1}^{N} \theta_j x_j}\right), \ldots, \left(\frac{i}{N}, \frac{\sum_{j=1}^{i} s(x_j)}{\sum_{j=1}^{N} \theta_j x_j}\right), \ldots, (1,1) \tag{A3}$$

Now

$$\frac{\sum_{j=1}^{i} s(x_j)}{\sum_{j=1}^{N} \theta_j x_j} = \frac{\sum_{j=1}^{N} s(x_j) - \sum_{j=i+1}^{N} s(x_j)}{\sum_{j=1}^{N} \theta_j x_j}$$

$$= \frac{\sum_{j=1}^{N} \theta_j x_j - \left[\sum_{j=i+1}^{N} \theta_j x_j \right]}{\sum_{j=1}^{N} \theta_j x_j}$$

So if the above number is larger than or equal to $\frac{\sum_{j=1}^{i} x_j}{T}$ for all i, we have shown that $L_{s(X)} \geq L_X$. We can exclude the case $x_{i+1}=\ldots=x_N=0$ since then the assertion is trivial. We have to show that

$$\frac{\sum_{j=1}^{i} \theta_j x_j + \sum_{j=i+1}^{N} \theta_j x_j}{\sum_{j=1}^{i} x_j + \sum_{j=i+1}^{N} x_j} = \frac{\sum_{j=1}^{N} \theta_j x_j}{T} \geq \frac{\left[\sum_{j=i+1}^{N} \theta_j x_j \right]}{\sum_{j=i+1}^{N} x_j}$$

Since for all $a,b,c,d \in \mathbb{R}^+$

$$\frac{a}{b} \geq \frac{c}{d} \Rightarrow \frac{a+c}{b+d} \geq \frac{c}{d} \geq \frac{[c]}{d}$$

it suffices to show that

$$\frac{\sum_{j=1}^{i} \theta_j x_j}{\sum_{j=1}^{i} x_j} \geq \frac{\sum_{j=i+1}^{N} \theta_j x_j}{\sum_{j=i+1}^{N} x_j} \qquad \text{(A4)}$$

From the lemma below we have that

$$\frac{\sum_{j=1}^{i}\theta_j x_j}{\sum_{j=1}^{i}x_j} = \alpha_i \in [\theta_i,\theta_1] \qquad \text{(A5)}$$

unless $x_1=\ldots=x_i=0$ but then, since X decreases, X is the zero-vector, which we exclude and (applied to $\theta_{i+1},\ldots,\theta_N,x_{i+1},\ldots,x_N$)

$$\frac{\sum_{j=i+1}^{N}\theta_j x_j}{\sum_{j=i+1}^{N}x_j} = \beta_i \in [\theta_N,\theta_{i+1}] \qquad \text{(A6)}$$

(since $x_{i+1}=\ldots=x_N=0$ is excluded).

Since $\theta_{i+1}\le\theta_i$, for all i, we have that (A5) and (A6) imply (A4), completing the proof of the theorem. □

Lemma :
$$\sum_{j=1}^{i}\theta_j x_j = \alpha_i\left(\sum_{j=1}^{i}x_j\right) \qquad \text{(A7)}$$

where $\alpha_i \in [\theta_i,\theta_1]$.

Proof : Since all $\theta_j \in [0,1]$ we have that

$$\sum_{j=1}^{i}\theta_j x_j \in \left[0,\sum_{j=1}^{i}x_j\right]$$

and hence

$$\sum_{j=1}^{i}\theta_j x_j = \alpha_i\left(\sum_{j=1}^{i}x_j\right) \qquad \text{(A8)}$$

for a certain $\alpha_i \in [0,1]$. Suppose now that $\alpha_i<\theta_i$. Hence $\alpha_i<\theta_j$ for all j=1,...,i and hence

$$\alpha_i\left(\sum_{j=1}^{i} x_j\right) = \sum_{j=1}^{i} \alpha_i x_j$$

$$< \sum_{j=1}^{i} \theta_j x_j$$

unless $x_1=\ldots=x_i=0$ in which case (A7) is trivial for any α_i. Hence (A9) contradicts (A8). In the same way, suppose $\alpha_i>\theta_1$. Hence $\alpha_i>\theta_j$ for all j=1,...,N, hence for all j=1,...,i. So

$$\alpha_i\left(\sum_{j=1}^{i} x_j\right) = \sum_{j=1}^{i} \alpha_i x_j$$

$$> \sum_{j=1}^{i} \theta_j x_j$$

contradicting (A8) again. Consequently, $\alpha_i \in [\theta_i,\theta_1]$. □

Appendix 2

Proof of Theorem III.1.1

We have to compare the Lorenz curves L_X of $X=(x_1,\ldots,x_N)$ and

$$s'(X) = (x_1,\ldots,x_{\frac{1}{\eta}-1},x_{\frac{1}{\eta}+1},\ldots,x_{\frac{2}{\eta}-1},x_{\frac{2}{\eta}+1},\ldots,x_{\frac{K-1}{\eta}+1},\ldots,x_{\frac{K}{\eta}-1},\underbrace{0,\ldots,0}_{K}) \qquad (A10)$$

To prove that $L_{s'(X)} \geq L_X$ we have to show (denote $M=\frac{1}{\eta}$) that

$$\frac{x_1+\ldots+x_{M-1}+x_{M+1}+\ldots+x_{2M-1}+\ldots+x_{iM-1}+x_{iM+1}+\ldots+x_{iM+j}}{T-\sum_{\ell=1}^{K}x_{\ell M}} \geq \frac{x_1+\ldots+x_{iM+j-i}}{T} \qquad (A11)$$

where i=1,...,K-1 ; j=1,...,M-1 (note that j=M-1 denotes the case where we have $x_{(i+1)M-1}$ as last term in the nominator of the left hand side and note that the case i=0 yields (A11) trivially, for all j=1,...,M-1). (A11) gives

$$x_1T+\ldots+x_{M-1}T+x_{M+1}T+\ldots+x_{2M-1}T+\ldots+x_{iM-1}T+x_{iM+1}T+\ldots+x_{iM+j}T$$
$$\geq x_1T+\ldots+x_{iM+j-i}T-x_M(x_1+\ldots+x_{iM+j-i})-x_{2M}(x_1+\ldots+x_{iM+j-i})-\ldots-x_{KM}(x_1+\ldots+x_{iM+j-i})$$

where KM = N.

In the right hand side, amongst the sum $x_1T+\ldots+x_{iM+j-i}T$ there are exactly

$$\alpha = \left[\frac{iM+j-i}{M}\right]$$

multiples of M, which remain after cancellation with the left hand side. This yields the equivalent condition

$$x_{iM+j-i+1}T + \ldots + x_{iM+j}T$$
$$\geq x_M T + x_{2M}T + \ldots + x_{\alpha M}T - x_M(x_1 + \ldots + x_{iM+j-i}) - x_{2M}(x_1 + \ldots + x_{iM+j-i}) - \ldots - x_{KM}(x_1 + \ldots + x_{iM+j-i})$$

Hence

$$x_{iM+j-i+1}T + \ldots + x_{iM+j}T + x_{(\alpha+1)M}(x_1 + \ldots + x_{iM+j-i}) + x_{(\alpha+2)M}(x_1 + \ldots + x_{iM+j-i}) + \ldots + x_{KM}(x_1 + \ldots + x_{iM+j-i})$$
$$\geq x_M(x_{iM+j-i+1} + \ldots + x_N) + x_{2M}(x_{iM+j-i+1} + \ldots + x_N) + \ldots + x_{\alpha M}(x_{iM+j-i+1} + \ldots + x_N)$$

$$= \left(\sum_{p=1}^{\alpha} x_{pM}\right)\left(\sum_{t=iM+j-i+1}^{iM+j} x_t + \sum_{s=iM+j+1}^{N} x_s\right) \qquad \text{(A12)}$$

Hence we have the condition

$$\left(\sum_{t=iM+j-i+1}^{iM+j} x_t\right)\left(T - \sum_{p=1}^{\alpha} x_{pM}\right) + \left(\sum_{q=\alpha+1}^{K} x_{qM}\right)\left(\sum_{r=1}^{iM+j-i} x_r\right) \geq \left(\sum_{p=1}^{\alpha} x_{pM}\right)\left(\sum_{s=iM+j+1}^{N} x_s\right) \qquad \text{(A13)}$$

Note that

$$T - \sum_{p=1}^{\alpha} x_{pM} \geq (M-1) \sum_{p=1}^{\alpha} x_{pM}$$

since X decreases, and, again since X decreases,

$$\sum_{t=iM+j-i+1}^{iM+j} x_t \geq \sum_{m=iM+j+1}^{iM+j+i} x_m \geq \sum_{m=iM+j+i+1}^{iM+j+2i} x_m \geq \ldots \geq \sum_{m=iM+j+(M-2)i+1}^{iM+j+(M-1)i} x_m$$

Hence

$$\left(\sum_{t=iM+j-i+1}^{iM+j} x_t\right)\left(T - \sum_{p=1}^{\alpha} x_{pM}\right)$$

$$\geq \left(\sum_{p=1}^{\alpha} X_{pM}\right)\left(\sum_{m=iM+j+1}^{iM+j+i} X_m + \sum_{m=iM+j+i+1}^{iM+j+2i} X_m + \ldots + \sum_{m=iM+j+(M-2)i+1}^{iM+j+(M-1)i} X_m\right)$$

(M-1 terms in the last factor)

$$= \left(\sum_{p=1}^{\alpha} X_{pM}\right)\left(\sum_{m=iM+j+1}^{iM+j+(M-1)i} X_m\right)$$

Hence it suffices to show, by (A13), that

$$\left(\sum_{p=1}^{\alpha} X_{pM}\right)\left(\sum_{m=iM+j+1}^{iM+j+(M-1)i} X_m\right)$$

$$+ \left(\sum_{q=\alpha+1}^{K} X_{qM}\right)\left(\sum_{r=1}^{iM+j-i} X_r\right) \geq \left(\sum_{p=1}^{\alpha} X_{pM}\right)\left(\sum_{s=iM+j+1}^{N} X_s\right)$$

which reduces to

$$\left(\sum_{q=\alpha+1}^{K} X_{qM}\right)\left(\sum_{r=1}^{iM+j-i} X_r\right) \geq \left(\sum_{p=1}^{\alpha} X_{pM}\right)\left(\sum_{s=iM+j+(M-1)i+1}^{N} X_s\right) \qquad \text{(A14)}$$

or

$$\frac{\sum_{p=1}^{\alpha} X_{pM}}{\sum_{q=\alpha+1}^{K} X_{qM}} \leq \frac{\sum_{r=1}^{iM+j-i} X_r}{\sum_{s=2Mi+j-i+1}^{N} X_s} \qquad \text{(A15)}$$

But for all p=1,...,α

$$x_{pM} \leq \frac{1}{M}(x_{(p-1)M+1} + \ldots + x_{pM})$$

and for all q=α+1,...,K-1

$$x_{qM} \geq \frac{1}{M}(x_{qM+1} + \ldots + x_{(q+1)M})$$

since X decreases, yielding, since N=KM,

$$\frac{\sum_{p=1}^{\alpha} x_{pM}}{\sum_{q=\alpha+1}^{K} x_{qM}} \leq \frac{\frac{1}{M}(x_1 + \ldots + x_{\alpha M})}{\frac{1}{M}(x_{(\alpha+1)M+1} + \ldots + x_N) + x_N}$$

$$\leq \frac{x_1 + \ldots + x_{\alpha M}}{x_{(\alpha+1)M+1} + \ldots + x_N}$$

$$\leq \frac{x_1 + \ldots + x_{iM+j-i}}{x_{(\alpha+1)M+1} + \ldots + x_N} \qquad \text{(A16)}$$

since αM=[iM+j-i] ≤ iM+j-i.

Furthermore

$$(\alpha + 1)\, M + 1$$

$$= \quad \alpha M + M + 1$$

$$\leq \quad iM + j - i + M + 1$$

$$\leq \quad 2iM + j - i + 1$$

since $i \geq 1$. Consequently (A16) gives

$$\frac{\sum_{p=1}^{\alpha} X_{pM}}{\sum_{q=\alpha+1}^{K} X_{qM}} \leq \frac{\sum_{r=1}^{iM+j-i} X_r}{\sum_{s=2iM+j-i+1}^{N} X_s}$$

which is (A15) and hence (A11). Consequently $L_{s'(X)} \geq L_X$ in all cases. □