A characterization of distributions which satisfy Price's Law and consequences for the Laws of Zipf and Mandelbrot

L. Egghe and R. Rousseau

Journal of Information Science 1986 12: 193
DOI: 10.1177/016555158601200406

A characterization of distributions which satisfy Price's Law and consequences for the Laws of Zipf and Mandelbrot

L. Egghe
LUC, Universitaire Campus, B-3610 Diepenbeek, Belgium; and UIA, Postbus 13, B-2610 Wilrijk, Belgium

R. Rousseau
KIH West-Vlaanderen, Zeedijk 101, B-8400 Oostende, Belgium; and, UIA, Postbus 13, B-2610 Wilrijk, Belgium

Received 3 July 1986

The distributions satisfying Price's Law are characterized. From this it is seen that the Laws of Zipf and of Mandelbrot can satisfy Price's Law only approximately. The form that Price's Law takes for these two distributions is then explicitly calculated. In the case of the Law of Mandelbrot, the same Law of Price is discovered as in "L. Egghe. An exact calculation of Price's Law for the Law of Lotka" (to appear in *Scientometrics*). This was expected, since the Laws of Lotka and of Mandelbrot are equivalent. The method used in this paper however is simpler.

1. Introduction

Suppose that we select a certain subject and that we look at all the researchers working in this area and that we also look at their publications. Suppose we have N researchers. Of course, as always, there are those publishing a lot of papers (the 'top-authors') and there are less prolific authors. Starting with the top-authors, the following question is posed. In order to find half of all publications, what geometrical fraction N^{α} ($\alpha \in]0, 1]$) of the authors do we need? Price's Law says that in many cases $\alpha \approx \frac{1}{2}$.

For instance, if we have $N = 100$ authors, then the $N^{1/2} = 10$ top-authors produce half of all the papers in this area. More generally, to find a fraction α of all publications, we need the top N^{α} authors.

This law was experimentally found to be approximately true for the so-called bibliometric laws: Zipf's Law, Mandelbrot's Law, and others. So far, exact calculations of Price's Law have been done only in [6] for the Law of Lotka, and in [7] where a bibliometric distribution is constructed, satisfying Price's Law for $\alpha = \frac{1}{2}$.

To be complete, we mention here also the possible study of the so-called (arithmetical) 80/20-rule: 20% of the top-authors produce 80% of the papers or, more generally: what fraction $100x$ % of the top-authors is needed in order to obtain 100θ % of the papers. This problem was studied in [4] for the classical bibliometric laws.

Of course, instead of 'authors-papers' one can also study 'journals-articles' or 'words-occurrences', and in fact all social phenomena.

In this paper we will characterize the distributions which satisfy exactly Price's Law. If we denote by $R(r)$ the cumulative number of papers of the authors on rank $1, \ldots, r$ (where we order the authors in decreasing order of productivity), we find that only situations where

$$R(r) = A \log r \tag{1}$$

is valid, satisfy Price's Law in an exact way. It is easy to see that all distributions satisfying (1) do indeed satisfy Price's Law (this was in fact also remarked by D. De Solla Price himself in [2]). We prove in this paper also the (easy) converse statement: if Price's Law is valid, then we have a situation which satisfies (1).

Since it is known that both the Laws of Zipf and Mandelbrot do not conform with (1) but approximate to it, it is then also logical to find an 'approximate' Price's Law for these distributions. This is calculated in the next two sections. Of course, since the Laws of Mandelbrot and of Lotka are formally equivalent (see e.g. [3]), and since we calculated already a generalization of Price's Law for the Law of Lotka (see [6]), we must rediscover this same generalized Law of Price

North-Holland
Journal of Information Science 12 (1986) 193–197

as valid for the Law of Mandelbrot. This is indeed the case. Furthermore, the calculations in our paper are easier than in [6], so that it is worth re-establishing this same Law of Price by our treatment.

To catch all types of sociological phenomena, we will use henceforth the term 'sources-items' to stand for 'authors-papers', 'journals-articles', 'words-occurrences' and 'employees-incomes', as a few examples.

2. Characterization of Price's Law

We prove, for any sociological distribution the following theorem.

Theorem. *Let there be N sources and denote by R(r) the cumulative number of items produced by the sources on rank* $1, \ldots, r$, *where we order the sources in decreasing order of productivity. Then the following assertions are equivalent*:

(i) $R(r) = A \log r$ *for every* $r = 1, 2, \ldots$, *where A is a constant.*

(ii) *The exact Law of Price is valid: for every* $N \in \mathbb{N}$ *and every* $\alpha \in]0, 1]$ *such that* $N^\alpha \in \mathbb{N}$, N^α *sources produce a fraction* α *of all items.*

Proof. (i) $\Rightarrow$ (ii). This is easy and well known (cf. [2]): we have, for every $\alpha \in]0, 1]$ and every $N \in \mathbb{N}$ such that $N^\alpha \in \mathbb{N}$:

$$\alpha R(N) = \alpha A \log N = A \log N^\alpha.$$

Hence, the top N^α sources produce a fraction α of all items (being $R(N)$).

(ii) $\Rightarrow$ (i). From (ii) we have

$$R(N^\alpha) = \alpha R(N) \tag{2}$$

for every $\alpha \in]0, 1]$ and $n \in \mathbb{N}$, such that $N^\alpha \in \mathbb{N}$, since $R(N)$ denotes the total number of items produced by all the sources. Since this is true for every $\alpha \in]0, 1]$, and $N \in \mathbb{N}$ such that $N^\alpha \in \mathbb{N}$ we can reach every rank $r = 1, 2, 3, \ldots$ by choosing, for every $r = 1, 2, 3, \ldots$ α so that $N^\alpha = r$. Hence $\alpha \log N = \log r$.

Thus (2) becomes:

$$R(r) = (\log r / \log N) R(N), \qquad R(r) = A \log r,$$

for every $r = 1, 2, 3, \ldots$ where $A = R(N)/\log N$ is a constant, for the given population of sources, independent of r. □

Corollary. *The Laws of Zipf and of Mandelbrot do not satisfy exactly Price's Law.*

Proof. (a) The Law of Zipf states that

$$f(r) = C/r,$$

where C is a constant and where $f(r)$ denotes the number of items produced by the source on rank r.

Hence

$$R(r) = \sum_{k=1}^{r} f(k) = \sum_{k=1}^{r} \frac{C}{k} \approx C(\log r + \gamma)$$

where $\gamma = 0.5772\ldots$ denotes Euler's number. Hence

$$R(r) = C \log r + D$$

with $D = C \cdot \gamma \neq 0$. This is not the form of $R(r)$ in the previous theorem.

(b) Mandelbrot's Law states that:

$$f(r) = A/(1 + Br),$$

where $f(r)$ is as above and where A and B are constants. As shown in [3], it follows that here

$$R(r) = (A/B) \log(1 + Br),$$

which is clearly not of the form of $R(r)$ in the previous theorem. □

Although Zipf's Law and the Law of Mandelbrot do not satisfy exactly Price's Law, the form of $R(r)$ in the previous corollary suggests that we might have an approximation of Price's Law. In the next two sections we will calculate the 'generalized' laws of Price for the Laws of Zipf, respectively Mandelbrot, i.e.: N^α sources produce a fraction θ of all items, and we investigate if and when $\alpha \approx \theta$.

3. The generalized Law of Price for the Law of Zipf

(A) To have a fraction θ (i.e. 100θ %) of all items, we find a number r of sources through the

condition

$$\sum_{k=1}^{r} f(k) = \theta \sum_{k=1}^{N} f(k)$$

or (using Zipf's Law: $f(k) = C/k$)

$$\sum_{k=1}^{r} \frac{C}{k} = \theta \sum_{k=1}^{N} \frac{C}{k}.$$

So, we find approximately

$$\log r + \gamma = \theta(\log N + \gamma)$$

or

$$\theta = \theta(r) = (\log r + \gamma)/(\log N + \gamma) \qquad (3)$$

or inversely

$$r = r(\theta) = \exp[\theta(\log N + \gamma) - \gamma].$$

Incidentally, by putting $x = r/N$, the fraction of all sources needed to have the fraction θ of all items, we find the arithmetic $100x/100\theta$-rule:

$$x = (1/N) \exp[\theta(\log N + \gamma) - \gamma].$$

For $\theta = 0.8$, we obtain the following values:

N	x
10	0.5 [1]
100	0.35
1000	0.22
10000	0.14
100000	0.09
1000000	0.06

For this result, the more exact formula

$$\sum_{k=1}^{r} \frac{1}{k} = \log r + \gamma + \frac{1}{2r}$$

is used: this formula is more accurate than the one used above, but is necessary only for small N. The above formula appears for instance in [8, p.2].

(B) From this result, we can derive the form of Price's Law where we have a Zipf Law. We look for α such that $N^{\alpha} = r$ sources yield 100θ % of the items. Hence $\alpha \log N = \log r$. So, by (3), we find

$$\alpha = \frac{\theta \log N - \gamma(1-\theta)}{\log N}. \qquad (4)$$

We see that for large N we have $\alpha \approx \theta$. Hence, in practice we have the exact law of Price. Some values are (for $\theta = 0.8$):

N	α
10	0.75
100	0.77
1000	0.78
10000	0.79
100000	0.79
1000000	0.79

So we conclude that even somewhat less than $N^{0.8}$ sources are needed to obtain a fraction 0.8 of all the items.

Remarks. (1) Formula (3) can also be expressed as a function of μ, the average number of items that a source produces. In the case of Zipf's Law, we have

$$\mu = \frac{1}{N} \sum_{r=1}^{N} f(r) = \frac{1}{N} \sum_{r=1}^{N} \frac{C}{r}.$$

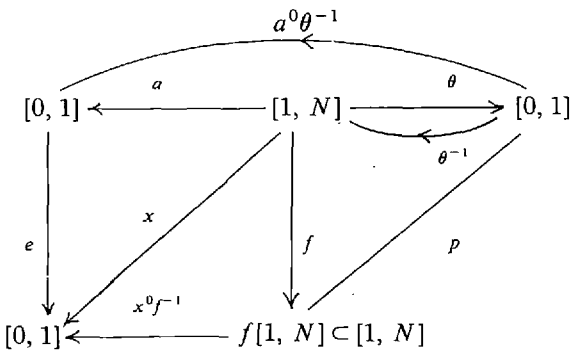
But we have $f(N) = 1$ (the source on the last rank produces 1 item). So $C = N$. Hence

$$\mu = \sum_{r=1}^{N} \frac{1}{r} \approx \log N + \gamma.$$

Hence (3) can be written as

$$\theta = (\log r + \gamma)/\mu.$$

(2) All the parameters described above behave as in the following commutative diagram:



$$f(r) = C/r \quad \text{(Zipf's Law)},$$
$$a(r) = \alpha = \ln r/\ln N,$$
$$e(\alpha) = N^{\alpha - 1}, \qquad x(r) = r/N,$$

$$p(f(r)) = \sum_{i=1}^{r} f(i) / \sum_{i=1}^{N} f(i),$$

$$\theta(r) = p(f(r)) \approx \frac{\ln r + \gamma}{\ln N + \gamma},$$

$$\theta^{-1}(t) = \exp(t\gamma + t \ln N - \gamma).$$

(C) *An application*. In [9] Ogden writes that though a dictionary of the English language may contain as many as 100 000 entries, or may run to upward of 500 000, there are not more than 60 000 words with which anyone but a specialist is likely to be concerned. But of these 60 000 at least 20 000 are of frequent occurrence in school textbooks or in reading matter regarded as suitable for the young. They form what is called BASIC English.

If we suppose that those 60 000 words obey Zipf's Law, how many of the word occurrences do we cover with the 20 000 words of BASIC English? So, in this case, where r and N are given, we need θ or α. Using (3) we find

$$\theta \approx \frac{\ln(20\,000) + \gamma}{\ln(60\,000) + \gamma} = 0.905.$$

So we may conclude that BASIC English covers 90% of the spoken language. We have here (using (4)) that $\alpha = 0.900$.

4. The generalized Law of Price for the Law of Mandelbrot

In this section we repeat the process used in the previous section, but now for the law

$$f(r) = A/(1 + Br)$$

We now have some additional difficulties: indeed, the form of $f(r)$ is more intricate to work with than in the Zipf case. But more fundamentally, contrary to the previous paragraph, where the constant C cancels in the calculations, the constant B here does not. So we need to know its value.

(A) *Calculation of the value of B*. We notice first of all that in the case where we have Mandelbrot's Law, we also have the (equivalent) Lotka's Law (cf. [3])

$$\phi(n) = D/n^2$$

where D is a constant and where $\phi(n)$ denotes the number of sources with n items. Following the same arguments as in [1] or in [6] (which is based upon practical situations) we have

$$\phi(n) = n_m^c/n^2$$

where $1 \leqslant c \leqslant 2$ and where n_m is the number of items in the most productive source (i.e. the source with rank 1). From this it follows (cf. also [3]):

$$r = \int_n^{n_m} \left(n_m^c/n'^2\right) \mathrm{d}n'$$

Hence, after an easy calculation,

$$f(r) = n = n_m \Big/ \left(\frac{1}{n_m^{c-1}} r + 1\right). \tag{5}$$

So $B = 1/n_m^{c-1}$. We note also the extreme cases.

$c = 1 : f(r) = n_m/(\mathrm{r} + 1)$: i.e. very close to Zipf's Law.

In this case we again expect $\theta \approx \alpha$.

$c = 2 : f(r) = n_m/((r/n_m) + 1)$: in this case we will establish α in function of θ and note that $\alpha \neq \theta$.

These results must also be identical with the results in [6], since there the generalized Law of Price has been calculated for Lotka's Law which is – as mentioned already – formally equivalent to Mendelbrot's Law.

(B) We now proceed as in the previous paragraph by putting

$$\sum_{k=1}^{r} \frac{A}{1 + Bk} = \theta \sum_{k=1}^{N} \frac{A}{1 + Bk}. \tag{6}$$

We now use

$$\sum_{k=1}^{x} \frac{1}{1 + Bk} \approx \int_1^x \frac{\mathrm{d}k}{1 + Bk} + D$$

with $0 < D < \gamma$. So, (6) becomes

$$\frac{1}{B} \log\left(\frac{1 + Br}{1 + B}\right) + D = \theta\left(D + \frac{1}{B} \log\left(\frac{1 + BN}{1 + B}\right)\right).$$

Hence

$$\log(1 + Br) = \theta(\log(1 + BN) - \log(1 + B) + DB) + \log(1 + B) - DB.$$

Since $B \leqslant 1$ and since $0 < D < \gamma$, we can write

$$\log(1 + Br) \approx \theta \log(1 + BN)$$

since N is large. So, in general

$$\theta = \theta(r) = \log(1 + Br)/\log(1 + BN). \quad (7)$$

Or conversely

$$r = r(\theta) = \frac{(1 + BN)^{\theta} - 1}{B}. \quad (8)$$

(C) From (8) we derive, as in the previous paragraph, the generalized Law of Price.

For $N^{\alpha} = r$, we have, using (8)

$$\alpha = \frac{\theta \log(1 + BN) - \log B}{\log N}. \quad (9)$$

But since $B = 1/n_m^{c-1}$ and since $N = n_m^2\pi^2/6$ (see [5, appendix] for the easy proof of this), we find

$$\alpha = \frac{\theta \log\left(1 + n_m^{3-c}\frac{\pi^2}{6}\right) + \log n_m^{c-1}}{\log\left(n_m^2\frac{\pi^2}{6}\right)}. \quad (10)$$

This is the generalized Law of Price for Mandelbrot's Law. Let us now examine the two extreme cases.

$B = 1$ (or, equivalently $c = 1$). In this case, we see from (9) or (10) that $\alpha \approx \theta$: we have almost exactly Price's Law.

$B = 1/n_m$ (or, equivalently $c = 2$). Here (10) implies

$$\alpha \approx \frac{\theta \log\left(n_m\frac{\pi^2}{6}\right) + \log n_m}{\log\left(n_m^2\frac{\pi^2}{6}\right)},$$

since $n_m\pi^2/6 \gg 1$. Also $\log \pi^2/6$ is negligible in comparison with $\log n_m$. So

$$\alpha \approx (\theta + 1)/2.$$

These two conclusions are exactly the same as in [6] where Price's Law was investigated for Lotka's Law. Since the Laws of Lotka and Mandelbrot are mathematically equivalent, we had to find the same generalized law of Price! This is indeed the case. It must however be pointed out that the reasoning made here to obtain this Law of Price is much simpler than in [6]: working with rank-order distributions (Zipf, Mandelbrot) is much easier in this context than working with frequency distributions (Lotka).

(D) We conclude with some practical values for the two extreme cases $c = 1$ and $c = 2$, both for $\theta = 0.8$.

$c = 1$

n_m	α
10	0.801
20	0.800
30	0.800
40	0.800
50	0.800

Hence $\alpha \approx \theta$ very accurately, even for small n_m.

$c = 2$

n_m	α
10	0.89
20	0.89
30	0.89
40	0.89
50	0.89

Hence $\alpha \approx (\theta + 1)/2 = 0.9$ very accurately, even for small n_m.

References

[1] P.D. Allison, D. de Solla Price, B.C. Griffith, M.J. Moravsik and J.A. Stewart, Lotka's Law: a problem in its interpretation and application, *Social Studies of Science 6* (1976) 269–276.

[2] D. de Solla Price, A general theory of bibliometric and other cumulative advantage processes, *J. Amer. Soc. Inf. Sci.* 27 (1976) 292–306.

[3] L. Egghe, Consequences of Lotka's Law for the Law of Bradford, *J. Documentation 41* (3) (1985) 173–189.

[4] L. Egghe, On the 80/20 rule, *Scientometrics 10* (1–2) (1986) 55–68.

[5] L. Egghe, Pratt's measure for some bibliometric distributions and its relation with the 80/20 rule, *J. Amer. Soc. Inf. Sci,* to appear.

[6] L. Egghe, An exact calculation of Price's Law for the Law of Lotka, *Scientometrics,* to appear.

[7] W. Glaenzel and A. Schubert, Price distribution. An exact formulation of Price's 'Square Root Law', *Scientometrics 7* (3–6) (1985) 211–219.

[8] I.S. Gradshtein and I.M. Ryzhik, *Tables of Integrals, Series and Products* (Academic Press, New York/London, 1965).

[9] C.K. Ogden, BASIC English, in: *Encyclopaedia Britannica* (1967).