Information Processing & Management Vol. 25, No. 5, pp. 527-544, 1989
Printed in Great Britain.

0306-4573/89 $3.00 + .00

ANALYSIS OF LOTKA'S LAW: THE SIMON-YULE APPROACH

YE-SHO CHEN
Department of Quantitative Business Analysis, Louisiana State University,
Baton Rouge, LA 70803, USA

(*Received* 12 *September* 1988, *accepted in final form* 19 *January* 1989)

Abstract—A major difficulty in using the well-known Lotka's law in information science is in the estimation of parameters. In this paper, we argue that the difficulty arises from the misuse of goodness-of-fit tests. As an alternative, we adopt Simon's five-step modeling process for the study of Lotka's law. Three significant contributions can be identified. First, an index approach is used to identify a general formulation of Lotka's law. Second, a time series approach is used to identify two influential variables associated with the empirical data. Third, the constructive mechanism proposed by Simon is used to derive a distribution resembling the general formulation of Lotka's law. Further research on refining the constructive mechanism is suggested.

1. INTRODUCTION

Building and testing a model of Lotka's law is a typical example of extreme hypotheses. According to Herbert A. Simon (Nobel Laureate), extreme hypotheses [1] are "assertions that a particular specific functional relation holds between the independent and the dependent variable." A standard practice for testing extreme hypotheses is the use of goodness-of-fit tests. Simon argues that those testing procedures are fundamentally unsatisfactory, since "an extreme hypothesis cannot be sensibly identified with the null hypothesis without shifting completely the burden of proof that is supposed to be assured by a new theory, and, what is worse, without making the tacit assumption that the correctness of a theory is an all-or-none matter and not simply a matter of goodness of approximation." (A different perspective of judging the plausibility of goodness-of-fit tests is discussed in Section 3.)

Instead, Simon [1] proposed a more constructive alternative to standard probabilistic and statistical test of fit. The modeling process is outlined below:

1. Begin with empirical data, not hypotheses.
2. Draw simple generalizations from striking features of the data.
3. Find limiting conditions by manipulating the influential variables associated with the data.
4. Construct simple mechanisms to explain the simple generalizations.
5. Propose the explanatory theories that go beyond the simple generalizations and make experiments for new empirical observations.

In this paper, we adopt and apply the five-step process to the modeling of Lotka's law. A brief review of Lotka's law is conducted in Section 2. In Section 4 we examine empirical data of Lotka's law by using the index approach proposed by Chen and Leimkuhler [3,8,9]. Some striking features of the data are observed and discussed in Section 5. In Section 6, we identify Lotka's law as a marginal property of a time series. Some influential variables affecting the property are also discussed. In Section 7, we focus on the generating mechanism proposed by Simon [10]. A significant contribution of the model is that it yields an equilibrium distribution resembling the observed phenomenon of Lotka's law. The need for successive refinements of the Simon-Yule model is discussed in Section 8. Finally, Section 9 presents the conclusion.

2. LOTKA'S LAW

In his well-known paper published in 1926, Lotka [2] examined patterns of scientific productivity among chemists. He discovered that if he classified his population of chemists according to how frequently they published, then the number of chemists publishing n papers, $f(n)$, was approximately equal to a/n^2, for some positive constant a, i.e.

$$f(n) = an^{-2}, \qquad n = 1,2,3,\ldots \tag{1}$$

Letting $F(n) = \sum_{i=n}^{\infty} F(i)$ be the number of authors contributing no less than n papers, then a frequently used alternative form of Lotka's law is

$$F(n) = \sum_{i=n}^{\infty} ai^{-2} \sim a\int_{n}^{\infty} \frac{1}{x^2}\,\mathrm{d}x = an^{-1}, \qquad n = 1,2,3,\ldots$$

This approximation is relatively accurate when n is not small.

Lotka's law of scientific productivity has been used to describe a wide variety of observation–class relationships [3], e.g. number of articles vs journals, and number of occurrences vs words. The law has also been suggested as being equivalent to the other two well-known empirical laws of information science [3]: Bradford's law of bibliographic scattering [4] and Zipf's law of word frequency [5].

Recently, a more general inverse-power form of Lotka's law,

$$f(n) = an^{-b}, \qquad b > 0,\ n = 1,2,\ldots \tag{2}$$

was investigated by Pao [6], in which a testing procedure for Lotka's law is proposed, including the estimation of a and b, and the goodness-of-fit test of the observed and computed frequency distributions. A minor modification of eqn (2) was later proposed by Nicholls [7]:

$$f(n) = an^{-b}, \qquad b > 0,\ n = 1,2,\ldots n_{\max}.$$

A testing procedure, modified after Pao, is also discussed. The recent research efforts provide us further insight into Lotka's law. However, several fundamental problems associated with the law are still unanswered and need to be examined carefully.

3. PROBLEMS OF USING GOODNESS-OF-FIT TESTS

Traditionally, four main steps are used in modeling Lotka's law [7]: (1) measurement and tabulation, (2) equation formulation; (3) parameter estimation; and (4) goodness-of-fit test. A goodness-of-fit procedure is a statistical test of a hypothesis that the sampled population is distributed in a specific way [11]. There are several statistics available for goodness-of-fit tests. Among those test statistics used, the chi-square test is probably the most common one. The crucial assumption underlying the chi-square test procedure is that the sample is a *random* sample [11], i.e. the observations are independently and identically distributed. However, the empirical data on Lotka's law fail to satisfy this assumption, since the data represent either a complete, or an incomplete set of a finite population.

The chi-square test also suffers from the problem of combining of classes. Coile [12] argued that this combining of classes is undesirable and suggested that the Kolmogrov–Smirnov one-sample test be used, believing it to be more powerful than the chi-square test. Several authors followed Coile's suggestion and used the Kolmogrov-Smirnov test as a goodness-of-fit tool [6,7]. Like the chi-square test, it seems to be a misuse to apply the Kolmogrov-Smirnov test to the data related to Lotka's law. First, the crucial assumption of the test is that the sample is a *random* sample. Second, the test is conservative (i.e. not exact), since the data are discrete, not continuous [11].

A formal study of the effect of dependency on conventional tests of fit is conducted by Gleser and Moore [13]. The significant contribution of their paper is that "confounding of positive dependence with lack of fit is a general phenomenon in the use of omnibus tests of fit." The finding suggests the use of traditional tests of fit for modeling Lotka's law as an inappropriate approach.

4. EXAMINING EMPIRICAL DATA: AN INDEX APPROACH

Two drawbacks are commonly associated with the equation formulation reported in the literature. First, they assume the independent variable n runs from one to infinity and, second, they assume that successive n's are consecutive without any "jump" or gaps. Typically the observed values of n, beyond the first smaller values, will "jump" to larger values in progressively larger steps; that is, they contain "gaps" and do not run consequently from one to infinity.

Realizing the first drawback some authors [7] limited the running of n from one to $n_{\max}$, the maximum value of n. However, the problem of jumps is still not addressed. Recently, a more realistic approach for modeling Lotka's law was proposed by Chen and Leimkuhler [3]. The main idea of the approach is to associate an index with each observed value of n. Several significant results have been obtained. First, a common functional relationship among Lotka's law, Bradford's law, and Zipf's law was derived [3]. Second, the droop phenomenon of Bradford's law was explained [8]; and third, the concave abnormality of Zipf's law was clarified [9]. We apply the index approach to Lotka's law in the following discussion.

We introduce the notion of an index $i = 1,2,\ldots m$, and let n_i denote the i^{th} different observed value of n so that $n_{i+1} > n_i$. Let $f(n_i)$ denote the number of authors having published n_i papers. The data in Table 1 are taken from Lotka's [2] chemistry data coming from an index of Chemical Abstracts.

The indices $i = 1,2,\ldots,m$, can be divided into three regions: where i is small, where i is close to m, and otherwise. For small i, usually $n_i = i$. For instance, in Table 1, $n_i = i$, for $i = 1,2,\ldots,34$. For i close to n, usually $f(n_i) = 1$. For example, in Table 1, $f(n_i) = 1$, for $i = 58,59,\ldots,66$. Let i_ℓ be the maximum i such $n_i = i$ and let i_u be the minimum i such that $f(n_i) \neq 1$ and $f(n_j) = 1$, $i \leq j \leq m$. Then we have the following three important properties:

$$n_i = i, \qquad 1 \leqq i \leqq i_\ell \tag{3}$$

$$n_i \cong i \text{ and } f(n_i) \cong 1, \qquad i_\ell + 1 \leqq i \leqq i_u - 1 \tag{4}$$

$$f(n_i) = 1, \qquad i_u \leqq i \leqq m. \tag{5}$$

Thus, the relationship between n_i and $f(n_i)$, $i = 1,2,\ldots n$, is governed by some known properties [eqns (3) through (5)] and unknown distributions $f(n_i)$, $1 \leqq i \leqq i_\ell$, and n_i, $i_u \leqq i \leqq m$. We summarize the relationships in Table 2.

Up to this point, we have not made any assumption on the distribution of $f(n_i)$, $i = 1,2,\ldots m$. However, by using the index approach, we can identify several inherent properties of $f(n_i)$, $i = 1,2,\ldots,m$. The next step is to investigate the unknown distributions $f(i)$, $1 \leqq i \leqq i_\ell$, and n_i, $i_u \leqq i \leqq m$ by posing some conditions.

5. SOME STRIKING FEATURES OF THE DATA

Let r_i denote the rank of authors contributing papers, so that $r_{i+1} > r_i$, $i = 1,2,\ldots,m-1$, and let $g(r_i)$ be the frequency of occurrence of papers published by the author(s)

Table 1. Lotka's data taken from *Chemical Abstracts* [2]

i	n_i	$f(n_i)$	i	n_i	$f(n_i)$
1	1	3991	34	34	4
2	2	1059	35	36	1
3	3	493	36	37	1
4	4	287	37	38	4
5	5	184	38	39	3
6	6	131	39	40	2
7	7	113	40	41	1
8	8	85	41	42	2
9	9	64	42	44	3
10	10	65	43	45	4
11	11	41	44	46	2
12	12	47	45	47	3
13	13	32	46	49	1
14	14	28	47	50	2
15	15	21	48	51	1
16	16	24	49	52	2
17	17	18	50	53	2
18	18	19	51	54	2
19	19	17	52	55	3
20	20	14	53	57	1
21	21	9	54	58	1
22	22	11	55	61	2
23	23	8	56	66	1
24	24	8	57	68	2
25	25	9	58	73	1
26	26	9	59	78	1
27	27	8	60	80	1
28	28	10	61	84	1
29	29	8	62	95	1
30	30	7	63	107	1
31	31	3	64	109	1
32	32	3	65	114	1
33	33	6	66	346	1

Table 2. Relationship between n_i and $f(n_i)$, $i = 1,2,\ldots,m$

Region	Range of indices	Known properties	Unknown distribution
I	$1 \leqq i \leqq i_\ell$	$n_i = i$	$f(i)$
II	$i_\ell + 1 \leqq i \leqq i_u - 1$	$n_i \cong i$ $f(n_i) \cong 1$	
III	$i_u \leqq i \leqq m$	$f(n_i) = 1$	n_i

with rank r_i. Also, let $F(n_i) = \sum_{k=i}^{m} f(n_k)$ be the number of authors having published no less than n_i papers. In terms of the index notations, one can derive the following relationships between r, n, F, and g [3,8,9]. For $i = 1,2,\ldots,m$,

$$r_i = F(n_{m-i+1}) \tag{6}$$

and

$$g(r_i) = n_{m-i+1}. \tag{7}$$

From the last two equations, we can prove the following theorem [9]:

Theorem 1.

For $i = 1,2,\ldots,m$,

$$g(r_i) = a(r_i + b)^{1/c} \tag{8}$$

$$\text{iff} \quad F(n_i) = dn_i^c - b \tag{9}$$

where a, b, c, and d are constants, and $a,d > 0$, $c < 0$, $da^c = 1$, and $b > -1$.

Equations (8) and (9), without the index notation, are general formulations of Zipf's law and Lotka's law, respectively [9]. So, Theorem 1 shows formally that the two laws are equivalent under some conditions. By taking one further step, we can describe the unknown distributions as shown in Table 2.

Corollary 1.

From Theorem 1, we have:

$$\text{(a)} \quad f(i) = d(i^c - (i+1)^c), \qquad 1 \leqq i \leqq i_\ell \tag{10}$$

$$\text{(b)} \quad n_i = a(m - i + 1 + b)^{1/c}, \qquad i_u \leqq i \leqq m \tag{11}$$

Proof. (a) For $1 \leqq i \leqq i_\ell$,

$$\begin{aligned} f(i) &= f(n_i) = F(n_i) - F(n_{i+1}) \\ &= F(i) - F(i+1) \\ &= d(i^c - (i+1)^c). \end{aligned}$$

(b) For $i_u \leqq i \leqq m$,

$$\begin{aligned} n_i &= g(r_{m-i+1}) = a(r_{m-i+1} + b)^{1/c} \\ &= a(m - i + 1 + b)^{1/c}. \end{aligned}$$

□

An immediate implication of Corollary 1 relates to the phenomena identified by Booth [14] concerning the occurrences of words of low frequency. Let D denote the number of different words used in a text, and $f(i)$ denote the number of words occurring i times, $i = 1,2,\ldots,5$. The first phenomenon reveals a remarkable consistency of the ratio $f(1)/D$ of single occurrences to the vocabulary in the text. The second phenomenon shows that for a given i, $i = 1,2,\ldots,5$, the ratio $f(i)/f(1)$ is approximately a constant for each of the texts Booth tested. The following corollary derives Booth's two findings. For more details, see Chen and Leimkuhler [15].

Corollary 2.

If $c = 1$, then

(a) $f(1)/D \cong 0.5$

(b) $f(i)/f(1) = \dfrac{2}{i(i+1)}, \qquad i = 1,2,3,4,5.$

Proof. (a) $D = F(1) = d - b$

$$f(1)/D = \frac{1 - 2^c}{1 - b/d} \cong 0.5, \qquad \text{if } c = 1.$$

(b) For $i = 1,2,3,4,5,$

$$f(i)/f(1) = \frac{i^c - (i+1)^c}{1 - 2^c} = \frac{2}{i(i+1)}, \qquad \text{if } c = 1. \qquad \square$$

Another important implication of Corollary 1 is shown in Table 3. There we see that regions I and III are governed by two simple hyperbolic functions. The function f in the first region models the distribution of the less productive authors. On the other hand, the function n in the third region models the papers' distribution of the highly productive authors. The finding implies that Lotka's law, or the related phenomena in information science, is the mixture of two different types of distributions, i.e. trivial-many and vital-few. To see how the distributions are generated, more exploration of empirical data is necessary. Without loss of generality, we use journal productivity data for our next discussion.

6. IDENTIFYING THE INFLUENTIAL VARIABLES: A TIME SERIES APPROACH

Figure 1 shows the historical growth of a bibliography [16] on bibliometrical distributions. The history shows that this topic did not receive much annual attention until the year 1969. From 1926, when Lotka's paper appeared, until 1969 there were never more than six papers in any one year. Seventeen papers appeared in 1969, which appears to have been a turning point in the history of the topic. From 1969 to 1982, 407 references have appeared at an annual average production rate of over 30 papers per year.

If we analyze the time series pattern for the year-by-year publication of the journals listed in the bibliography, more information will show up. A partial list of the time series is shown in Table 4, where journals publishing more than four papers are considered. This table shows that the two most productive journals in this field have a long and continuing history of publication. They are the *Journal of the American Society of Information Science* and the *Journal of Documentation*. The years 1969 and 1975 were highly productive for the *Journal of Documentation*. The papers in 1969 were concerned largely with the theoretical foundations of the field, while those in 1975 concentrated on applications. The papers by Leimkuhler [17] and Brookes [18] appear to have had a significant influence on the burst of publications in 1969, which marks a turning point in the growth of the literature.

One interesting observation is available immediately from the time-series presentation of publication pattern. It shows that the data set $f(n_i)$ vs n_i, $i = 1,2,\ldots,m$, is obtained from the total column on the right-hand side of the representation. Thus, the empirical phenomena shown in Theorem 1 and Table 3 are marginal properties of the time series. Two variables might be influential for the generation of the time series, the entrance of *new* journals and the productivity of *old* journals. As we can see from Table 4, in every given year there is a chance for a new journal to publish a paper in the field. On the other hand, the productivity of an old journal at a certain point of time is roughly proportional to its previous publication. This is consistent with the concept of "richer gets richer."

Table 3. Relationship between n_i and $f(n_i)$, $i = 1,2,\ldots,m$ *deriving from Theorem* 1

Region	Range of indices	Known properties	Properties derived from Theorem 1
I	$1 \leqq i \leqq i_\ell$	$n_i = i$	$f(i) = d(i^c - (i+1)^c)$
II	$i_\ell + 1 \leqq i \leqq i_u - 1$	$n_i \cong i$ $f(n_i) \cong 1$	
III	$i_u \leqq i \leqq m$	$f(n_i) = 1$	$n_i = a(m - i + 1 + b)^{1/c}$

```
1926   1   *                                     <---------- Lotka [2]
1927   0
1928   0
1929   1   *
1930   0
1931   0
1932   1   *
1933   0                                         <---------- Bradford [4]
1934   1   *
1935   1   *
1936   0
1937   0
1938   2   **
1939   0
1940   0
1941   3   ***
1942   0
1943   1   *
1944   2   **
1945   0
1946   0
1947   0
1948   2   **                                    <--------- Zipf [5]
1949   1   *
1950   1   *
1951   0
1952   1   *
1953   4   ****
1954   2   **
1955   1   *
1956   4   ****
1957   2   **
1958   2   **
1959   2   **
1960   3   ***
1961   3   ***
1962   4   ****
1963   4   ****
1964   3   ***
1965   3   ***
1966   3   ***
1967   6   ******
1968   2   **
1969   17  *****************
1970   12  ************
1971   9   *********
1972   26  **************************
1973   27  ***************************
1974   29  *****************************
1975   37  *************************************
1976   46  **********************************************
1977   36  ************************************
1978   34  **********************************
1979   40  ****************************************
1980   61  *************************************************************
1981   39  ***************************************
1982   4   ****
```

Fig. 1. The yearly pattern of the bibliography on bibliometrical distributions [16].

7. THE CONSTRUCTIVE MECHANISM PROPOSED BY SIMON

The generating mechanism proposed by Simon [10] for skew distributions incorporates the concept of new and old entities. In terms of scientific publications, Simon's generating process can be stated in the following assumptions, where $f(n, t)$ is the number of different authors who have published exactly n papers in the first t papers of the time series.

Table 4. The time series pattern of the journals publishing more than four papers [16]

	Year																													
Journal	41	48	52	53	55	56	57	58	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	Total
1													1	1	1	1	1		2	3	3	5	3	5	7	4	4	3	2	46
2		1	1					1	1					1	1	9		2	3	1	1	8	3	2	2	2	3	3		45
3																				1	2	2	1		1	1	1	1		10
4																									1		6	2		9
5																				2	1	2				2				8
6																				1			1					6		8
7							1		1		1										1		1		1		1			7
8																2	3	1												6
9																							1	1		1		3		6
10																						1		2		2				5
11																		1				2	1				1			5
12																								3		1		1		5
13									1											1	2					2				6
14																							2	2	1					5
15																							3			1	1			5
16																						1	1	1	1					4
17										1										1						1	1			4
18				1	1	1		1																						4
19	1											1							1						1					4
Total	1	1	1	1	1	1	1	2	3	1	1	1	1	2	2	12	5	4	6	10	10	21	17	16	15	17	18	19	2	192

> Assumption I: There is a constant probability, α, that the $(t + 1)$-st paper be published by a new author—an author who has not published in the first t papers.

> Assumption II: The probability that the $(t + 1)$-st paper is written by an author who has published n papers is proportional to $nf(n,t)$—that is to the total number of papers written by all the authors who have published exactly n papers.

Letting v be the total number of different authors contributing papers under our consideration, the *expected* number of authors having published n papers derived from the two assumptions is

$$f_\theta(n) = v\rho B(n,\rho + 1), \qquad n = 1,2,3,\ldots, \tag{12}$$

where $\rho = 1/(1 - \alpha)$ and $B(n,\rho + 1)$ is the beta function with parameters n and $\rho + 1$. Simon calls the last equation a Yule distribution because Yule's paper [19], which predated the modern theory of stochastic processes, derived the same equation in a study of a biological problem. A formal proof of eqn (12) is shown in Appendix A.

Let $f(n)$ be the actual number of authors having published a count of n papers. If n is not in the index set $\{n_1,n_2,\ldots,n_m\}$, $f(n) = 0$. Let $e(n)$ be the deviation between the actual number and expected number of authors publishing n papers, then

$$f(n) = f_\theta(n) + e(n), \qquad n = 1,2,3,\ldots, \tag{13}$$

and

$$F(n) = F_\theta(n) + \epsilon(n), \qquad n = 1,2,3,\ldots, \tag{14}$$

where

$$F(n) = \sum_{k=n}^{\infty} f(n), \quad F_\theta(n) = \sum_{k=n}^{\infty} f_\theta(n), \text{ and } \epsilon(n) = \sum_{k=n}^{\infty} e(k).$$

We show in the following theorem that eqn (9) in Theorem 1 can be derived from eqn (12) with some minor difference. Thus, Simon's generating mechanism provides a theoretical foundation for the phenomena identified in Section 5. Note that $f(n)$ denotes the same things in eqns (1) and (13).

THEOREM 2.
For $i = 1,2,\ldots,m$,

$$F(n_i) = v\rho\Gamma(\rho)n_i^{-\rho} + (O(n_i^{-\rho-1}) + \epsilon(n_i)). \qquad (15)$$

Proof. From Ijiri and Simon [1], we have

$$F_\theta(n) = v\ \rho\beta(n,\rho), \qquad n = 1,2,3,\ldots \qquad (16)$$

The beta function can be rewritten [20] as follows:

$$B(n,\rho) = \Gamma(\rho)n^{-\rho} + O(n^{-\rho-1}), \qquad (17)$$

where the O notation means that there are positive constants c and n_0 such that for n equal to or greater than n_0, the third term of eqn (17) is less than or equal to $cn^{-\rho^{-1}}$. By substituting eqns (16) and (17) into eqn (14), we have

$$F(n) = v\rho\Gamma(\rho)n^{-\rho} + O(n^{-\rho-1}) + \epsilon(n), \qquad n = 1,2,3,\ldots \qquad (18)$$

Since with real data, we are only interested in the index set $\{n_1,n_2,\ldots,n_m\}$, we rewrite eqn (18) as eqn (15). □

If we set $d = v\rho\Gamma(\rho)$, $c = -\rho$, and $b_i = -(O(n_i^{-\rho-1}) + \epsilon(n_i))$, when b_i, $i = 1,2,\ldots,m$, are approximately equal, then we can replace b_i by b (or define b as the average of all b_i's) and rewrite eqn (15) as eqn (9).

8. NEED FOR FURTHER REFINEMENTS

Up to this point, we have attempted to provide a sound generating mechanism to explain the general form of Lotka's law shown in eqn (9). The resulting distribution, eqn (15), shows us some explanatory capability of the model. However, there is room for further improvement since we have no knowledge of the residuals $\epsilon(n_i)$ and $e(n_i)$ for $i = 1,2,\ldots,m$.

One way to analyze the residuals is to plot them. Plots of residuals are useful in at least three aspects [21]: (1) assessing the adequacy of the fitted model; (2) uncovering a dependence of the data on factors missing in the models; and (3) pointing to other improvements in the model. Two such plots and their uses are described below. In general, these plots should appear as a scatter of points with no pattern or structure. Patterns are characteristic of residuals from incomplete models, that is, models that can be improved.

The first plot is the plot of residuals against the variable included in the model. This plot gives useful information about the dependence of the response on the dependent variable. When this dependence has been approximately formulated, this plot will appear as a random scatter. Trends or patterns suggest a more complex dependence. The second plot is the plot of residuals against fitted values. If a good model has been approximately fitted to the data, the residuals tend to resemble random noise and exhibit no marked dependence on other variables. Thus, a plot of residuals against the fitted values should appear as patternless random scatter.

To illustrate the usefulness of these two plots, let us consider the data listed in Simon [10]. The data contained the scientific publications in *Chemical Abstracts* over 10 years, in a history of physics, and in *Econometrica* over a 20-year period. Table 5 gives the actual data, the estimate values, and the residuals. (Note that eqn (13) is used to make the

Table 5. Number of persons contributing (data available from Simon [10])

No. of Contributions	*Chemical Abstracts*			Physicists			*Econometrica*		
	Actual	Estimate	Residual	Actual	Estimate	Residual	Actual	Estimate	Residual
1	3,991	4,050	−59	784	824	−40	436	453	−17
2	1,059	1,160	−101	204	217	−13	107	119	−12
3	493	552	−29	127	94	33	61	51	10
4	287	288	−1	50	50	0	40	27	13
5	184	179	5	33	30	3	14	16	−2
6	131	120	11	28	20	8	23	11	12
7	113	86	27	19	14	5	6	7	−1
8	85	64	21	19	10	9	11	5	6
9	64	49	15	6	8	−2	1	4	−3
10	65	38	27	7	6	1	0	3	−3
11 or more	419	335	84	48	52	−4	22	25	−3
Estimated ρ		1.43			1.64			1.69	

estimates in Table 5.) Figures 2 and 3 give the plots of residuals in *Chemical Abstracts*. Figures 4 and 5 give the plots in Physicists. Figures 6 and 7 give plots in *Econometrica*. Visually, we see that each of Figs. 3, 5, and 7 consists of four clusters. On the other hand, Fig. 2 represents an up-going trend, Figs. 4 and 6 reveal a similar pattern where the first three points increase steadily and then go up and down alternately. All the figures suggest the incompleteness of the model used. The source of the trends and clusters is an open question that needs more study.

Our next step is to examine the two basic assumptions described in Section 7 and to look for an additional explanatory variable that could be incorporated into the generating mechanism so as to lead to a better second approximation of the empirical data. The process of successive refinements was conducted by Simon and his colleagues from 1955 to 1977. Instead of doing research on scientific publications, they focused on the sizes of

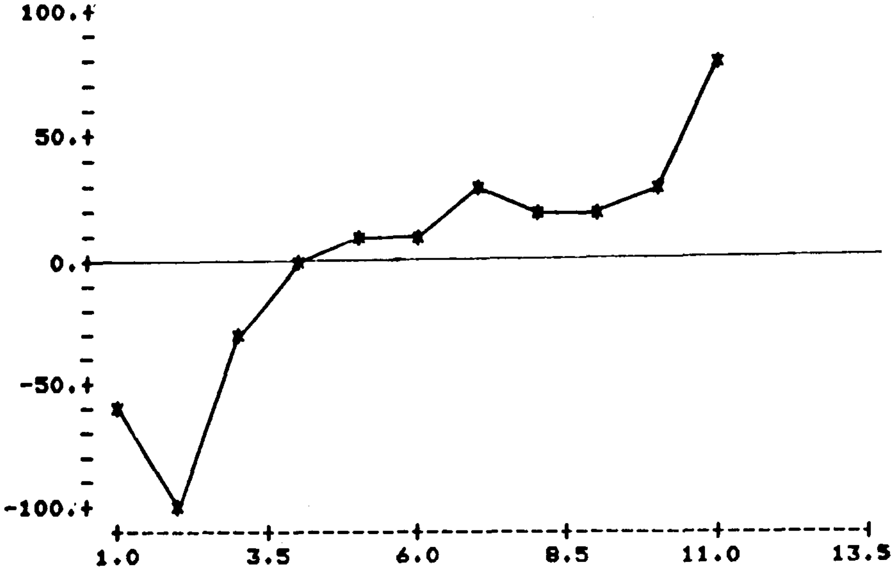

Fig. 2. The plot of residuals against the number of contributions in *Chemical Abstracts*.

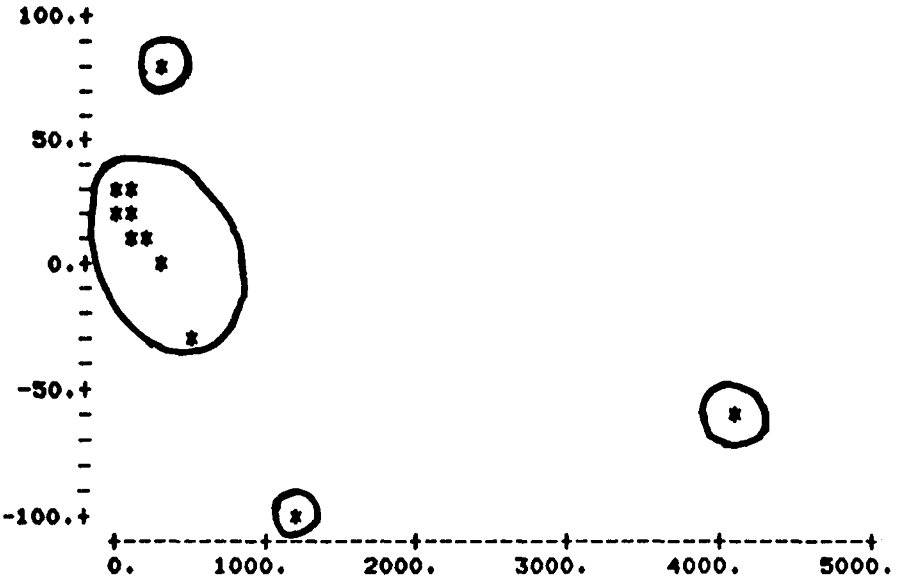

Fig. 3. The plot of residuals against fitted values in *Chemical Abstracts*.

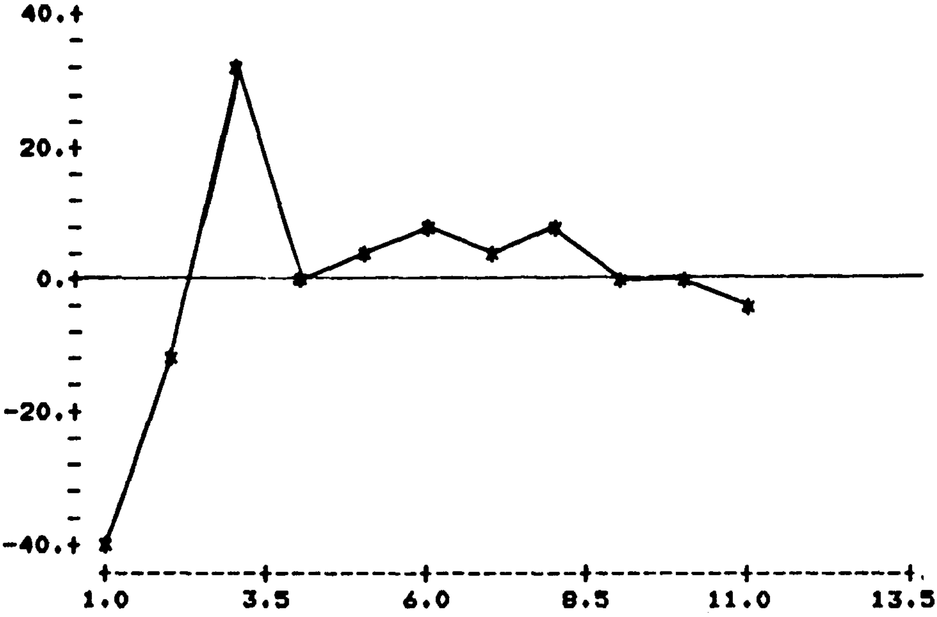

Fig. 4. The plot of residuals against the number of contributions in Physicists.

business firms. Eleven papers were collected in the monograph: Skew Distribution and the Sizes of Business Firms [1].

According to Simon, the refined assumptions were based on empirical data and supported by economical theory. The two main refinements are: (1) autocorrelation in form growth rates, which gives a significant modification of the law of proportionate effect, is incorporated in the generating mechanism; and (2) the effects of mergers, acquisitions, and dissolutions upon concentration and the sizes of business firms are examined. The significant results of the two refinements are that they provide two different economic explanations for the concavity of the bilogarithmic firm-size distributions as observed in empirical data.

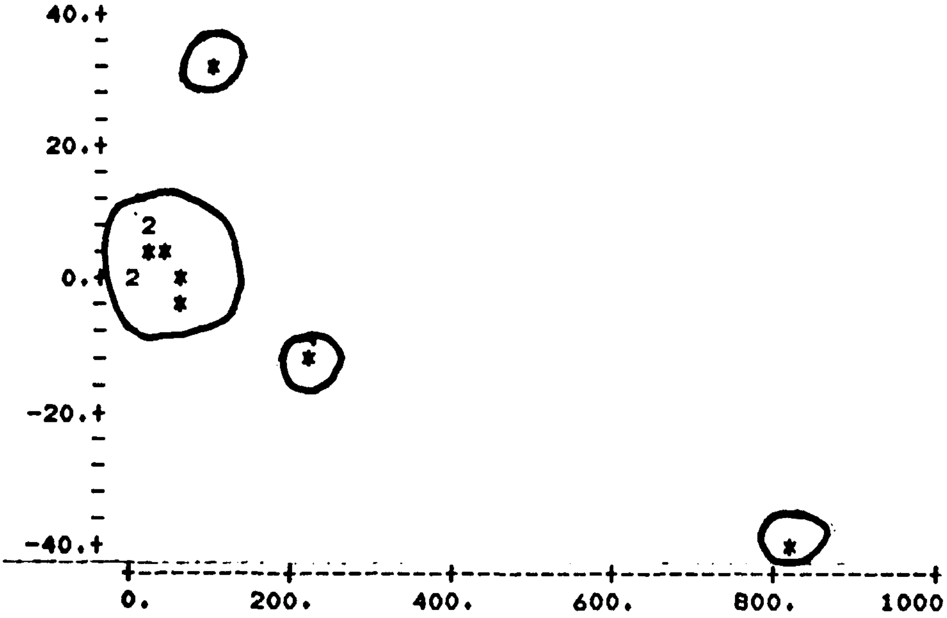

Fig. 5. The plot of residuals against fitted values in Physicists.

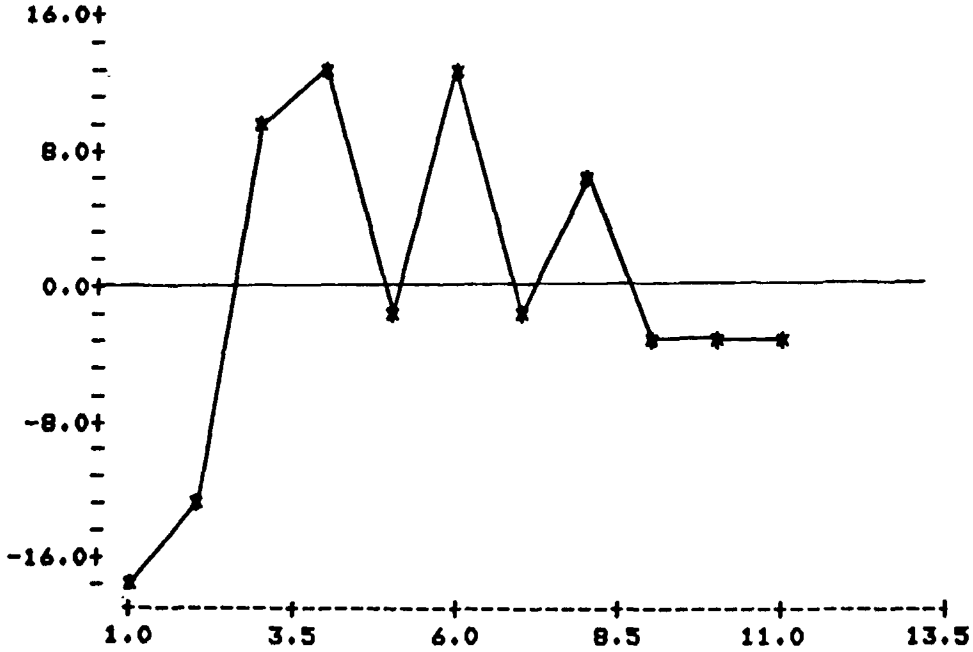

Fig. 6. The plot of residuals against the number of contributions in *Econometrica*.

What all of this implies is that it is logical to follow the accumulated knowledge developed for the sizes of business firms and apply it to scientific publications to provide a better and richer understanding of Lotka's law. As an example, consider Simon's refinement assumptions on business mergers and acquisitions. A possible relevance of the assumptions to Lotka's law is the hypotheses about how the forming and dissolution of joint authors could affect author output. Empirical data are necessary to support the hypotheses.

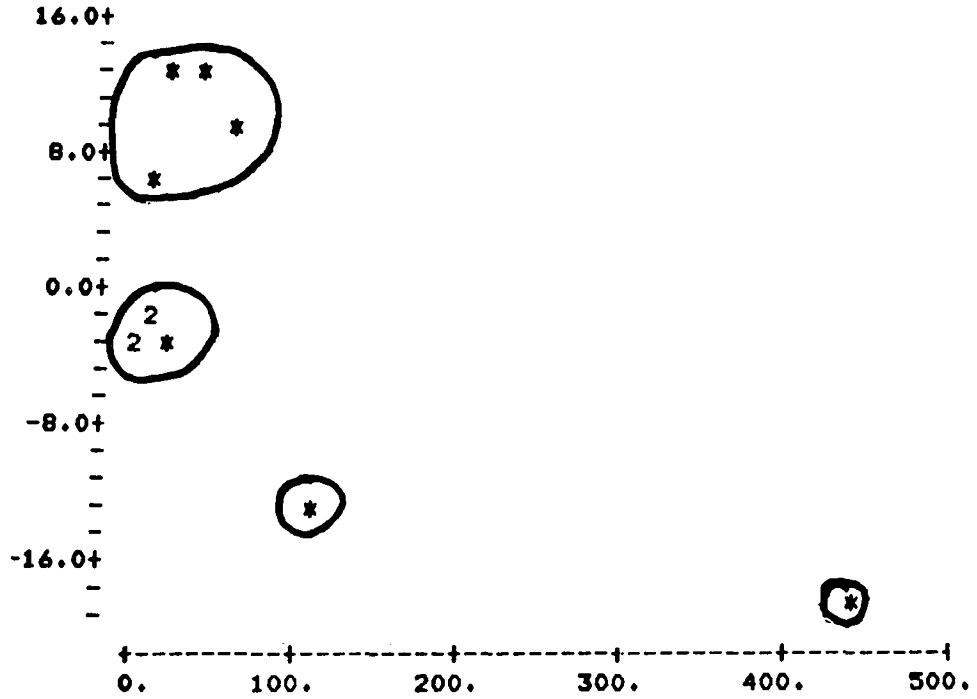

Fig. 7. The plot of residuals against fitted values in *Econometrica*.

9. CONCLUSION

Three significant approaches for studying Lotka's law are discussed in this article. First, an index approach is applied to identify a general formulation of Lotka's law. Several implications of the new formulation are discussed. Second, a time series approach is used to identify two influential variables associated with the empirical data. Third, a generating mechanism incorporating the two influential variables is used to derive an equilibrium distribution resembling the general formulation of Lotka's law. The constructive mechanism was proposed and successively modified by Simon for studying the sizes of business firms. The process of successive refinements is suggested for the study of Lotka's law.

Acknowledgement—This research was supported by the College of Business Administration at Louisiana State University. The authors also thank the referees for their valuable comments.

REFERENCES

1. Ijiri, Y.; Simon, H.A. Skew distributions and the sizes of business firms. New York: North-Holland Publishing Company; 1977.
2. Lotka, A.J. The frequency of distribution of scientific productivity. Journal of the Washington Academy of Science, 16: 317-323; 1926.
3. Chen, Y.S.; Leimkuhler, F.F. A relationship between Lotka's law, Bradford's law, and Zipf's law. Journal of the American Society for Information Science, 37: 307-314; 1986.
4. Bradford, S.C. Sources of information on specific subjects. Engineering, 137: 85-86; 1934.
5. Zipf, G.K. Human behavior and the principle of least effort. Reading, MA: Addison-Wesley; 1949.
6. Pao, M.L. Lotka's law: A testing procedure. Information Processing and Management, 21: 305-320; 1985.
7. Nicholls, P.T. Empirical validation of Lotka's law. Information Processing & Management, 22: 417-419; 1986.
8. Chen, Y.S.; Leimkuhler, F.F. Bradford law: An index approach. Scientometrics, 11: 183-198; 1987.
9. Chen, Y.S.; Leimkuhler, F.F. Analysis of Zipf's law: An index approach. Information Processing and Management, 23: 171-182; 1987.
10. Simon, H.A. On a class of skew distribution function. Biometrika, 52: 425-446; 1955.
11. Conover, W.J. Practical nonparametric statistics. Second ed. New York: John Wiley & Sons; 1980.
12. Coile, R.C. Lotka's frequency distribution of scientific productivity. Journal of the American Society for Information Science, 28: 166-370; 1977.

13. Gleser, C.J.; Moore, D.S. The effect of dependence on chi-squared and empiric distribution tests of fit. The Annals of Statistics, 11: 1100–1108; 1983.
14. Booth, A.D. A law of occurrences for words of low frequency. Information and Control, 10: 386-393; 1967.
15. Chen, Y.S.; Leimkuhler, F.F. Booth's law of word frequency. Journal of the American Society for Information Science, In press.
16. Leimkuhler, F.F.; Chen, Y.S.; Johnson, B.D. Analysis and application of information productivity models. 1-5 Progress Report, School of Industrial Engineering, Purdue University; NSF Grant IST-70118933A1; March 1983.
17. Leimkuhler, F.F. The Bradford distribution. Journal of Documentation, 23: 199-207: 1967.
18. Brookes, B.C. The derivation and application of the Bradford-Zipf distribution. Journal of Documentation, 24: 247-265; 1968.
19. Yule, G.U. A mathematical theory of evolution, based on the conclusions of Dr. J.C. Willis, F.R.S. Philosophical Transactions of the Royal Society of London, Series B, 213: 21-87; 1924.
20. Titchmarsh, E.C. The theory of functions. Second ed. Oxford: Clarendon Press; 1939.
21. Dantzig, D.V. Data analysis, exploratory. International Encyclopedia of Statistics. In: W.H. Kruskal and J.M. Tanur, editors. New York: Macmillan and Free Press; 1978.
22. Chen, Y.S.; Leimkuhler, F.F. A type-token identity in the Simon-Yule model of text. Journal of the American Society for Information Science, 40: 45-53; 1989.
23. Chen, Y.S. An exponential recurrence distribution in the Simon-Yule model of text. Cybernetics and Systems, 19: 521-545; 1988.
24. Haight, F.A.; Jones, R.B. A probabilistic treatment of qualitative data with special reference to word association tests. Journal of Mathematical Psychology, 11: 237-244; 1974.

APPENDIX A: PROOF OF EQN (12)

Simon's generating process was recently applied to examine a type-token identity associated with the plays of Shakespeare [22] and an exponential recurrence phenomenon in written texts [23]. No doubt, Simon's model is outstanding and has been cited frequently. However, the mathematical derivations in his paper are not rigorous. Simon himself described them as "heuristic." Haight and Jones [24] argued that "the proof given by Simon is a little cloudy" and they reproved eqn (12) by using the probability-generating function technique. In the following we clarify Simon's proof by providing more formal derivations. Note that Simon's model was originally proposed for the modeling of text generation. For the purpose of comparison, we adopt his terminologies in the following discussion.

Consider a book that is being written, and has reached a length of t words. Let $f(n,t)$ be the number of different words that have occurred exactly n times in the first t words. For example, if there are 300 different words that have occurred exactly once each, then $f(1,t) = 300$.

LEMMA A.1.

Given $f(n,t)$, then for $t = 1,2,\ldots,$ and $n = 2,3,\ldots t$, we have

$$f(n,t+1) = \begin{cases} f(n,t) - 1, \textit{ with prob.} = 1 - \alpha \dfrac{nf(n,t)}{t}; \\ f(n,t), \textit{ prob.} = \alpha + (1-\alpha)\left(1 - \dfrac{nf(n,t)}{t} - \dfrac{(n-1)f(n-1,t)}{t}\right); \\ f(n,t) + 1, \textit{ with prob.} = (1-\alpha)\dfrac{(n-1)f(n-1,t)}{t}. \end{cases}$$

Also, given $f(1,t)$, then for $t = 1,2,\ldots,$ we have

$$f(1,t+1) = \begin{cases} f(1,t) - 1, \textit{ with prob.} = (1-\alpha)\dfrac{f(1,t)}{t}; \\ f(1,t), \textit{ with prob.} = (1-\alpha)\left(1 - \dfrac{f(1,t)}{t}\right); \\ f(1,t) + 1, \textit{ with prob.} = \alpha \end{cases}$$

Proof. Given $f(n,t)$, the introduction of the $(t+1)$st word will result in one of the following conditions:

1. $f(n,t+1) = f(n,t) - 1$, i.e. one of the words that has occurred exactly n times in the first t words is used. The probability for this event is

$$(1-\alpha)\frac{nf(n,t)}{t}$$

2. $f(n,t+1) = f(n,t) + 1$, i.e. one of the words which has occurred exactly $n-1$ times in the first t words is used. The probability for this event is

$$(1-\alpha)\frac{(n-1)f(n-1,t)}{t}.$$

3. $f(n,t+1) = f(n,t)$, i.e. the introduction of the $(t+1)$st word does not change the frequency of $f(n,t)$. The probability for this event is

$$\alpha + (1-\alpha)\left(1 - \frac{nf(n,t)}{t} - \frac{(n-1)f(n-1,t)}{t}\right).$$

Similar tokens are for $f(1,t)$, $t = 1,2,\ldots$ □

The last lemma plays an important role in the following derivations.

Lemma A.2.
For $t = 1,2,\ldots,$ *and* $n = 2,3,\ldots t$, *we have*

$$E(f(n,t)) = \frac{1-\alpha}{t}\left[(n-1)E(f(n-1,t)) - nE(f(n,t))\right]$$

and

$$E(f(1,t+1) - E(f(1,t)) = \alpha - \frac{1-\alpha}{t}E(f(1,t)).$$

Proof. From Lemma A.1, we have, for $t = 1,2,\ldots,$ and $n = 2,3,\ldots,t$,

$$\begin{aligned}
E(f(n,t+1)|f(n,t)) = {} & [f(n,t)-1]\left[(1-\alpha)\frac{nf(n,t)}{t}\right] \\
& + f(n,t)\left[\alpha + (1-\alpha)\left(1 - \frac{nf(n,t)}{t} - (n-1)\frac{f(n-1,t)}{t}\right)\right] \\
& + [f(n,t)+1]\left[(1-\alpha)\frac{(n-1)f(n-1,t)}{t}\right] \\
& = f(n,t)\cdot\alpha + f(n,t)\cdot(1-\alpha) \\
& \quad - (1-\alpha)\frac{nf(n,t)}{t} + (1-\alpha)\frac{(n-1)f(n-1,t)}{t} \\
& = f(n,t) - (1-\alpha)\frac{nf(n,t)}{t} + (1-\alpha)\frac{(n-1)f(n-1,t)}{t}
\end{aligned}$$

iff

$$E(f(n,t+1)|f(n,t)) - f(n,t) = \left(\frac{1-\alpha}{t}\right)[(n-1)f(n-1,t) - nf(n,t)].$$

This implies

$$\begin{aligned}&E(E(f(n,t+1)|f(n,t))) - E(f(n,t))\\ &\quad= \left(\frac{1-\alpha}{t}\right)[(n-1)E(f(n-1,t)) - nE(f(n,t))].\end{aligned}$$

Thus

$$\begin{aligned}&E(f(n,t+1)) - E(f(n,t))\\ &\quad= \frac{1-\alpha}{t}\,[(n-1)E(f(n-1,t) - nE(f(n,t))].\end{aligned}$$

On the other hand, for $t = 1,2,\ldots$

$$\begin{aligned}&E(f(1,t+1)|f(1,t))\\ &\quad= [f(1,t)-1]\left[(1-\alpha)\left(1-\frac{f(1,t)}{t}\right)\right]\\ &\qquad+ f(1,t)\left[(1-\alpha)\left(1-\frac{f(1,t)}{t}\right)\right]\\ &\qquad+ [f(1,t)+1]\alpha\\ &\quad= f(1,t)\cdot\alpha + \alpha + f(1,t)\cdot(1-\alpha) - (1-\alpha)\frac{f(1,t)}{t}\end{aligned}$$

iff

$$E(f(1,t)|f(1,t)) - f(1,t) = \alpha - \left(\frac{1-\alpha}{t}\right)f(1,t).$$

This implies

$$E(E(f(1,t+1)|f(1,t))) - E(f(1,t)) = \alpha - \left(\frac{1-\alpha}{t}\right)E(f(1,t)).$$

Thus

$$E(f(1,t+1)) - E(f(1,t)) = \alpha - \left(\frac{1-\alpha}{t}\right)E(f(1,t)). \qquad \square$$

Simon assumed that there is a steady-state solution in which the expected frequencies of all classes of words change in the same proportion. To be more specific, the steady-state assumption can be stated as follows:

Definition A.1. The random process $f(n,t)$ reaches a steady state if and only if

$$\frac{E(f(n,t+1))}{E(f(n,t))} = \frac{t+1}{t},$$

where $t = 1,2,\ldots,$ and $n = 1,2,\ldots,t$.

An alternative form of Simon's steady-state assumption is

$$\frac{E(f(n,t))}{t\alpha} = \frac{E(f(n,t+1))}{(t+1)\alpha} = h(n),$$

where $t = 1,2,\ldots$ and $n = 1,2,\ldots,t$. That is all expected frequencies will grow proportionally with t, so that they will maintain the same relative size. We denote the *expected relative frequency* as $h(n)$, $n = 1,2,\ldots,t$. Note that $n\alpha$ is the expected number of different words used in the text. Simon's steady-state assumption looks reasonable if we consider a text being "warmed up" for a fairly long time before the actual frequency count of the used words starts.

THEOREM A.1.

Under the steady-state assumption, we can derive from the Simon–Yule model

$$h(n) = (1+\rho)B(n,\rho+1)h(1), \quad \rho = \frac{1}{1-\alpha}, \quad h(1) = \frac{1}{2-\alpha}, \; n = 2,3,\ldots,t$$

Proof. From Lemma A.2, we have

$$E(f(n,t+1)) - E(f(n,t)) = \frac{1-\alpha}{t}\,[(n-1)E(f(n-1,t)) - nE(f(n,t))]$$

and

$$E(f(1,t-1)) - E(f(1,t)) = \alpha - \left(\frac{1-\alpha}{t}\right)(E(f(1,t)),$$

where $t = 1,2,\ldots,$ and $n = 1,2,\ldots,t$. When in the steady state, the equations become

$$h(n)(t+1) - h(n)t = \frac{1-\alpha}{t}\,[(n-1)th(n-1) - nth(n)]$$

and

$$h(1)(t+1) - h(1)t = 1 - \frac{1-\alpha}{t}\,th(1),$$

where $t = 1,2,\ldots,$ and $n = 2,\ldots,t$. These imply

$$h(1) = \frac{1}{2-\alpha}$$

and

$$h(n) = (1-\alpha)(n-1)h(n-1) - (1-\alpha)nh(n)$$

iff

$$(1 + (1 - \alpha)n)h(n) = (1 - \alpha)(n - 1)h(n - 1).$$

The last equation indicates that $h(n)$ is not equal to zero, for $n = 2,3,\ldots t$. Therefore

$$\begin{aligned} h(n) &= \frac{(1 - \alpha)(n - 1)}{1 + (1 - \alpha)n} h(n - 1) \\ &= \frac{n - 1}{n + \rho} h(n - 1), \qquad \text{where } \rho = \frac{1}{1 - \alpha} \end{aligned}$$

iff

$$h(n) = \frac{n - 1}{n + \rho} h(n - 1)$$

$$h(n - 1) = \frac{n - 2}{(n - 1) + \rho} h(n - 2)$$

$$h(n - 2) = \frac{n - 3}{(n - 2) + \rho} h(n - 3).$$

Thus,

$$\begin{aligned} h(n) &= \frac{n - 1}{n + \rho} \cdot \frac{n - 2}{(n - 1) + \rho} \cdots \frac{2 - 1}{2 + \rho} h(1) \\ &= \frac{(n - 1)!\,[(2 + \rho - 1)(2 + \rho - 2)\Gamma(\rho)]}{[(n + \rho)\ldots(2 + \rho)]\,[(2 + \rho - 1)(2 + \rho - 2)\Gamma(\rho)]} h(1) \\ &= \frac{\Gamma(n)\Gamma(\rho + 2)}{\Gamma(n + \rho + 1)} h(1) \\ &= \frac{(\rho + 1)\Gamma(n)\Gamma(\rho + 1)}{\Gamma(n + \rho + 1)} h(1) \\ &= (1 + \rho)\beta(n,\rho + 1)h(1) \\ &= (1 + \rho)\beta(n,\rho + 1)h(1), \end{aligned}$$

where

$$h(1) = \frac{1}{2 - \alpha}, \qquad n = 2,3,\ldots,t \qquad \square$$

The expected frequency, $f_\theta(n)$, $n = 1,2,\ldots$, is then $f_\theta(n) = vh(n) = v\rho B(n,\rho + 1)$.