

【予稿集】

標本サイズの増大に伴って形状が変わる Lotka 型確率密度関数の導出

"Success breeds success" (SBS) 確率過程に基づく解

小野寺 夏生

科学技術・学術政策研究所

nt.onodera@y5.dion.ne.jp

"Success breeds success"(SBS)理論を用いて逆冪型の科学生産性分布 (Lotka 型分布) を説明するいくつかのモデルが提案されてきた。しかしそれらの多くは定常分布を仮定しているため、標本サイズが増加するときの分布形の変化を説明できず、また、ある次数以上のモーメントが発散するという問題点を含む。この研究の目的は、モーメントが発散しない非定常的な Lotka 型分布を導くためのモデルを提出し、このモデルに沿って確率密度関数を導くことである。

Derivation of a non-stationary Lotka-type distribution A solution based on the "Success breeds success" (SBS) stochastic process

Natsuo ONODERA*

The National Institute of Science and Technology Policy (NISTEP)

1. はじめに

計量情報学の分野において、source が生産または所有する item の数に関する分布を一般的に科学生産性分布と呼ぶ。この分布を、item 数 j を持つ source の相対頻度 (確率分布) $f(j)$ で表す。最もよく知られている科学生産性分布である Lotka の法則は次式で表される。

$$f(j) = \frac{C}{j^a} \quad (j=1, 2, \dots) \quad (a > 1) \quad (1)$$

Lotka 自身は、Chemical Abstracts や他の書誌で観測した著者の生産する論文数の分布から Lotka の法則(1)を見出した[1]。この他、文章中の単語の出現頻度[2]、雑誌に掲載されるある主題の論文数[3]、論文が受ける引用回数[4,5]等の分布が近似的にこの法則に従う。これらを一般化すると、source (著者、単語等) が生産あるいは保有する item (論文、単語出現等) の数の分布と行うことができる。多くの経験的な事例では、(1)における a は 1.5~3 程度の範囲にわたっている[6]。

Lotka 型分布を導くため種々の確率モデルが提

出されているが、最も著名なのは"Success breeds success" (SBS)理論に基礎を置いたものである。しかしこれらのモデルには2つの問題点がある。

第一に、ほとんどのモデルが定常状態を扱っていることである。一般に、item 数(N)が増えるにつれて確率分布 $f(j)$ の形は変化する (すなわち分布が非定常的)。経験的にこの変化を詳しく分析した研究はある[7]が、これまでの Lotka 型モデルの研究ではこのことが説明されていない。

第二に、定常的分布ではある次数以上のモーメントが定義できない。Lotka 分布(1)の場合、 a より小さい最大の整数を m とするとき、 m 次以上のモーメントは発散する。理論的確率密度関数が発散すると、実測分布を理論分布に当てはめる方法が制限される等の不都合が起こる：

この研究では、モーメントが発散しない Lotka 型分布を導くためのモデルを提出し、確率密度関数を導く。この関数は系に含まれる items 数 N あるいは source 数 M をパラメータとし、それらの変化に伴う分布の変化を知ることができる。

2. 過去の SBS モデル研究の概観

Simon[8]に始まる SBS モデルの基本的考えは、「総計 N 個の item が生産された後、 $(N+1)$ 個目の item を、既に j 個の item を持つ source のいずれかが生産する確率は、それらの source が既に生産した item 数に比例する」である。

しかし、この仮定だけでは新しい source が出現することはできない。item の蓄積につれて source の数も増加する（例えば新しい著者の出現やテキスト中への新しい語の出現）ようにするため、多くの SBS モデルは、 $(N+1)$ 個目の item を新しい source が生産する「新 source 出現確率」（これを $\alpha(N)$ とする）の仮定を導入する。

このモデルに基づいて次の漸化微分方程式が導かれる（ $N \gg 1$ とする）。

$$\frac{\partial n_s(j)}{\partial N} = \begin{cases} \alpha(N) - \frac{1 - \alpha(N)}{N} n_s(1) & \text{for } j = 1 \\ \frac{1 - \alpha(N)}{N} \{(j-1)n_s(j-1) - (j)n_s(j)\} & \text{for } j \geq 2 \end{cases} \quad (2)$$

$n_N(j)$ は、既に N 個の item が蓄積されているとき、item 数 j を持つ source の数である。

また、 $\alpha(N)$ に対する上記の定義から次のように書ける（やはり $N \gg 1$ とする）。

$$\alpha(N) = \frac{dM}{dN} \quad (3)$$

これまでに提案された主な SBS モデルとそれから導かれた確率分布を表 1 に要約する。

表 1 これまでに提案された主な SBS モデル

提案者	$\Pr(j \rightarrow j+1)$	α	分布の N 依存性	得られる分布性
Simon, 1955 [8]	$\propto j \quad (j \geq 1)$	一定	なし (定常的)	Yule 分布 ($\kappa > 1$)
Simon, 1955 [8]	$\propto j \quad (j \geq 1)$	N とともにゆるやかに減少	なし (定常的)	Yule 分布 (N に依存する κ)
Egghe and Rousseau, 1995 [9]	$\propto j \quad (j \geq 1)$	事前の仮定なし	準定常的	Yule 分布 (N に依存する κ)
Schubert and Glänzel, 1984 [10]	$\propto (j + \beta) \quad (j \geq 0)$	事前の仮定なし	あり (但し定常部分に注目)	Waring 分布+時間依存項
Yule, 1925 [11]	$\propto j \quad (j \geq 1)$	事前の仮定なし	あり	Yule 分布+時間依存項

3. SBS 理論に基づく厳密な非定常的確率分布

SBS モデルに基づくこれまでの研究のほとんどは定常状態における分布に注目している。Yule[11]は非定常分布についてもかなり深く考察しているが、定式化には至っていない。この節では、 $\alpha(N)$ に対する仮定を設け、(2)に基づいて厳密な非定常的確率密度関数 $f_N(j)$ を導く。

3.1 基本的な方程式

分布を得るための基本方程式は(2)であるが、数学的に取り扱いやすいという理由により、以降では N の代わりに M を独立変数とする。(3)から、

$\frac{d}{dN} \rightarrow \frac{d}{dM} / \frac{dN}{dM} = \alpha(M) \frac{d}{dM}$ 。故に(2)は次のように書き換えられる。

$$\frac{dn_s(j)}{dM} = \begin{cases} 1 - \frac{1}{N} \left(\frac{1}{\alpha(M)} - 1 \right) n_s(1) & \text{for } j = 1 \\ \frac{1}{N} \left(\frac{1}{\alpha(M)} - 1 \right) \{(j-1)n_s(j-1) - j n_s(j)\} & \text{for } j \geq 2 \end{cases} \quad (4)$$

$\alpha(M)$ の関数形が与えられれば、(4)を解いて source 数 M における $n_M(j)$ を得ることができる。発表者は、4つの $\alpha(M)$ の仮定に対して非定常的確率密度関数を得たが、ここでは、最も妥当と考えられる仮定に対する結果を述べる。

3.2 確率密度関数

$\alpha(M)$ に対し以下を仮定する。

$$\alpha(M) = \frac{1}{1+(\mu/\kappa)} \quad (5)$$

ここで、 $\mu=N/M$, $\kappa>0$

(5)を用いると(4)は次式になる。

$$\frac{dn_M(j)}{dM} = \begin{cases} 1 - \frac{1}{\kappa M} n_M(j) & \text{for } j=1 \\ \frac{1}{\kappa M} \{(j-1)n_M(j-1) - jn_M(j)\} & \text{for } j \geq 2 \end{cases} \quad (6)$$

初期状態 $M=1$ において $n(1)=1$, $n(j)=0$ ($j \geq 2$) としてこの微分方程式を解くと以下の解が得られる。

$$n_M(j) = \frac{1}{y} \left\{ \kappa \int_0^1 t^{j-1} (1-t)^{\kappa-1} dt - \kappa \int_0^1 t^{j-1} (1-t)^{\kappa-1} dt + y^{\kappa-1} (1-y)^{\kappa-1} \right\} \quad (7)$$

$$= \frac{1}{y} \left\{ \kappa B(\kappa+1, j) - \kappa B(\kappa+1, j) + y^{\kappa-1} (1-y)^{\kappa-1} \right\} \quad (j \geq 1)$$

$$y = 1/M^{1/\kappa} \quad (8)$$

$B(\cdot, \cdot)$ はベータ関数、 $B_y(\cdot, \cdot)$ は不完全ベータ関数である。(7)と(8)を用いて確率密度関数 $f_y(j)$ は、

$$f_y(j) = \frac{n_M(j)}{M} = \kappa B(\kappa+1, j) - \kappa B(\kappa+1, j) + y^{\kappa-1} (1-y)^{\kappa-1} \quad (j \geq 1) \quad (9)$$

この確率分布は、 M の変化により形状が変わる非定常分布であり、任意のモーメントはある有限値に収束する。

3.3 確率密度関数の性質

3.3.1 M の変化に伴う分布形の変化

(9)の $f_y(j)$ は3つの項からなる。これらを f_1 、 f_2 、 f_3 とする。 f_1 は定常項 (M によらない) で $M \rightarrow \infty$ の場合に当たる。定性的には f_2 はこれを押し下げ、 f_3 は逆に押し上げる。このことを確かめるために、 $\kappa=1.0$ として $M=50$, 5000 のそれぞれの場合に対し、次の3式の j による変化を図1に示した：① f_1 ($M \rightarrow \infty$ での定常項)、② f_1+f_3 、③ $f_1+f_3+f_2$ ($f_y(j)$ に等しい)

図1から、 $M=50$ と $M=5000$ の間で③すなわち $f_y(j)$ の形が異なることが解る。

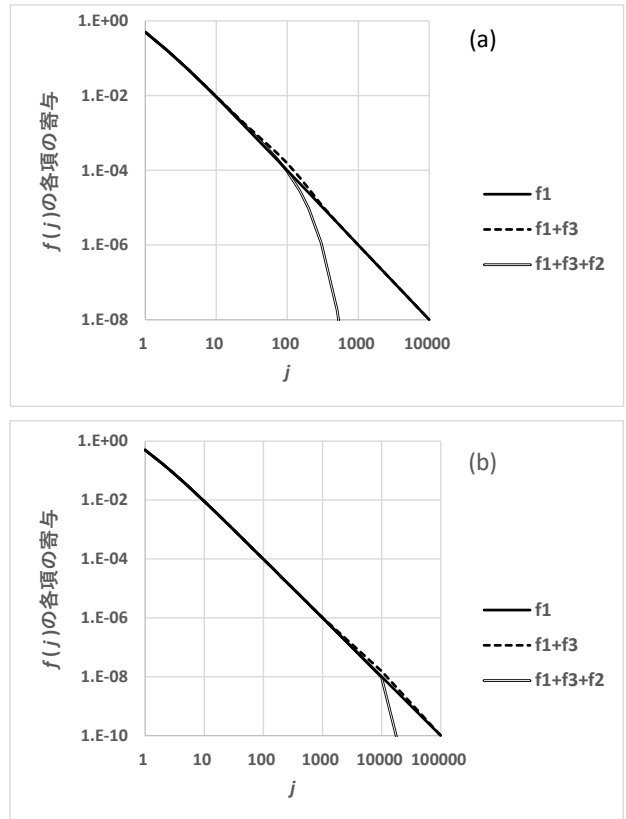


図1 $f_y(j)$ の各項の寄与 ($\kappa=1$ の場合) :
(a) $M=50$, (b) $M=5000$

①と②の比較から、第3項の押し上げ効果は極めて僅かなことが解る。これに対して、②と③の比較から解る第2項の押し下げ効果は非常に大きい。 j の大きい領域では第2項は第1項とほぼ同じ大きさ(符号は逆)となり、両者を合わせた $f_y(j)$ は急激に減衰する。

$M \rightarrow \infty$ ($y \rightarrow 0$) では(9)の第2項と第3項が消えて第1項のみが残り、

$$f_1(j) = \lim_{y \rightarrow 0} f_y(j) = \kappa B(\kappa+1, j) = \kappa \frac{\Gamma(\kappa+1)\Gamma(j)}{\Gamma(j+\kappa+1)} \quad (10)$$

という Yule 分布になる。この解は、Simon(1955)による定常分布を仮定した解[8] (表1参照) に等しい。(10)は $j \rightarrow \infty$ で $a=\kappa+1$ の Lotka の法則(1)に近づく。

3.3.2 M による分布の平均値及び分散の変化

(9)から分布の平均 μ と分散 σ^2 が得られるが、数式は省略し、 κ が 0.5、1.0、1.5、2.0、2.5 の5つの場合における M による変化を図2に示す。

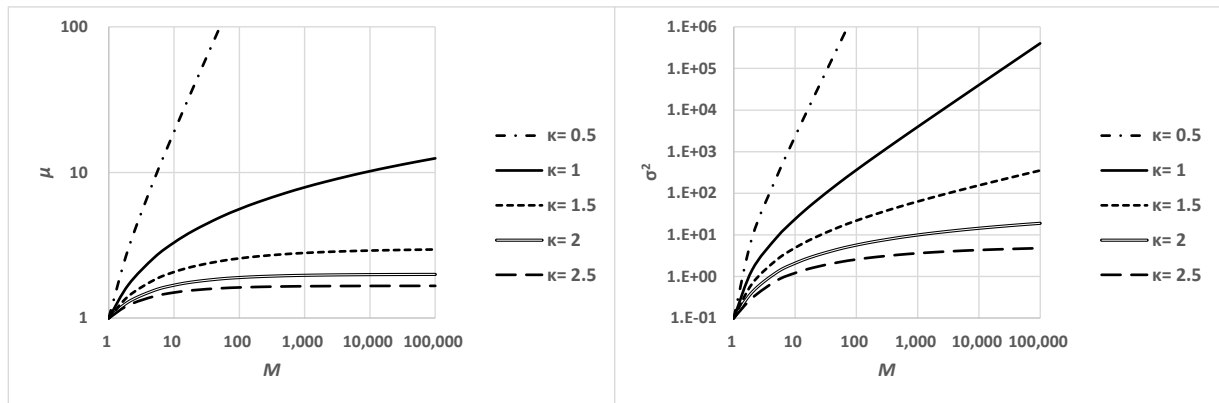


図2 source数(M)に伴う平均(μ) (左) と分散(σ^2) (右) の変化

$M \rightarrow \infty$ の極限を考えると、 μ は $\kappa > 1$ では $\kappa/(\kappa-1)$ に収束するが $0 < \kappa \leq 1$ では発散する。 σ^2 は $\kappa > 2$ では $\kappa^2/[(\kappa-1)^2(\kappa-2)]$ に収束するが $0 < \kappa \leq 2$ では発散する。

4. 結論

本発表では、定常分布の条件を課さない確率過程方程式を解くことにより、非定常的でモーメントの発散しない確率密度関数 $f(x)$ を得た。この分布は、 $N \rightarrow \infty (M \rightarrow \infty)$ の極限では定常的なYule分布になる。また、過去において非定常部分を含めて考察した唯一の例であるYuleのモデル[11]の非定常部分(定式化はされていない)が、3.で得た分布と等価であることを示すことができる。

注・文献

[1] Lotka, A. J. (1926). The frequency distribution of scientific productivity. *J Washington Academy Sci*, 16, 317-323.
 [2] Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Cambridge, MA: Addison Wesley.
 [3] Bradford, S. C. (1948). *Documentation*. London: Crosby Lockwood.
 [4] Price, D. S. (1976). A general theory of bibliometric and other cumulative advantage processes. *J Am So Info Sci*, 27(5/6), 292-306.

[5] Seglen, P. O. (1992). The skewness of science. *J Am So Info Sci*, 43(9), 628-638.
 [6] Haitun, S. D. (1982). Stationary scientometric distributions. II. Non-Gaussian nature of scientific activities. *Scientometrics*, 4(2), 89-104.
 [7] Oluic-Vukovic, V. (1998). Simon's generating mechanism: Consequences and their correspondence to empirical facts. *J Am So Info Sci*, 49(11), 867-880.
 [8] Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika*, 42(3/4), 425-440.
 [9] Egghe, L., & Rousseau, R. (1995). Generalized Success-Breeds-Success principle leading to time-dependent information distributions. *J Am So Info Sci*, 46(6), 426-445.
 [10] Schubert, A., & Glänzel, W. (1984). A dynamic look at a class of skew distributions. A model with scientometric applications. *Scientometrics*, 6(3), 149-167.
 [11] Yule, G. U. (1925). A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F. R. S.. *Philosoph Trans Royal Soc London, Series B*, (213), 21-87.