

Zipf-Bradford 分布

——計量情報学における基本的モデル——

小野寺 夏生

1. はじめに

ある事象の生起回数 x に対する確率分布関数として最もよく知られているのは、各試行の独立性に基礎を置く二項分布や Poisson 分布であろう。しかし、情報科学を含む社会科学の分野では、これらの偶然型確率分布よりも裾の広がった形の、逆べき型分布 ($\propto x^{-\rho}$) が観測されることが多い。次のような例が昔からよく知られている。

- ① 一定の長さの文章の中に x 回出現する語彙の相対度数 (Zipf [10])
- ② ある主題に関して一定期間内に x 件の論文を掲載する雑誌の相対度数 (Bradford [1])
- ③ ある期間に x 件の論文を発表した研究者の相対度数 (Lotka [4])
- ④ 人口が x 万人台の都市の相対度数 (Zipf [10])
- ⑤ 年収が x 百万円台の世帯の相対度数 (Pareto [7])
- ⑥ 事業規模 (売上げ高, 研究開発資金等) が x 億円台の企業の相対度数 (Ijiri & Simon [3])
- ⑦ x 種類の「種」を含む生物の「属」の相対度数 (Willis [9])

この種の分布を表現する式には、伝統的にいくつかの変形があるが、それらは結局等価な表現な

おのでら なつお 日本科学技術情報センター技術管理室 〒100 千代田区永田町 2-5-2

ので、本稿では一括して Zipf-Bradford 分布と呼ぶ。この事情について 2. で述べる。そして、この型の分布が、各試行における生起確率が独立ではない (過去の試行結果に依存する) ことを前提とした確率モデルから導かれることを、偶然型の Poisson 分布と対比しつつ 3. ~ 4. で説明する。最後に 5. では、従来の Zipf-Bradford 分布の式と多くの観測例のあいだに指摘されていた不一致がこのモデルにより解釈可能となることを示す。

なお、本稿で使用する記号とその意味を表 1 にまとめた。

2. Zipf-Bradford 分布関数

2.1 いくつかの表現の変形

a) 生起回数一存在確率 (Lotka 型)

生起回数 x 回の個体の存在確率 $f(x)$ が、

$$(1) f(x) \propto x^{-\rho}$$

にしたがう。 ρ は経験的に 2 前後の数である。

b) ランク一生起回数 (Zipf 型)

個体を生起回数の高い順に並べた時、 r 番目の個体の生起回数 $x(r)$ が次式にしたがう。

$$(2) x(r) \propto (r+r_0)^{-\sigma}$$

Zipf 自身の式は $\sigma=1, r_0=0$ であったが、その後経験的に上記の式に拡張されている。

c) ランク一累積生起回数 (Bradford 型)

ランク r までの累積生起回数 $S(r)$ が、

$$(3) S(r) \propto \ln\{1+(r/r_0)\}$$

で表わされる。

表 1 本稿で使用する記号の意味

記 号	説 明	導入された式の番号
$x, x(r)$	ある個体(箱)がもつ生産性(生起頻度). $x=0, 1, 2, \dots$	1, 2
$r, r(x)$	個体(箱)を生産性の大きい順に並べたときのランク. $r=1, 2, \dots, M$	2, 31
x_m	生産性最大, すなわち $r=1$ の個体(箱)のもつ生産性.	31, 32
M	個体(箱)の総数.	7, 8
N	全個体の生産性の総和. 生産性をボールの個数で表わすとき, ボールの総数.	7
$p_N(x)$	ボール数 N の時点で, $(N+1)$ 個目のボールが生産性 x の箱に入る確率.	4, 9, 11, 14
α	$p_N(x)$ の中に現われるパラメータで, 確率過程の偶然性を示す尺度.	14
θ	$\theta=1+\alpha M$. 確率分布関数に現われるパラメータ.	17
ν	$\nu=\alpha N$. 確率分布関数に現われるパラメータ.	18
$f_N(x), f_\nu(x)$	生産性の総和(ボールの総数)が N の時点において, 生産性(ボールの収容個数)が x である個体(箱)の存在確率. 本稿の中心である確率分布関数である.	4, 19
$f_{\nu \gg 1}(x), f_{\nu \ll 1}(x)$	$\nu \gg 1$ および $\nu \ll 1$ の各極限における漸近的な $f_\nu(x)$.	27, 29
ρ	$f_N(x) \propto x^{-\rho}$ の形で近似される場合の逆べき指数.	1, 12, 30
μ	$\mu=\nu/(1+\theta)$. $f_{\nu \gg 1}(x)$ の中に現われるパラメータ.	28
$A_1(x)$	$f_\nu(x)$ の解を ν で展開したときの係数.	21
$S(r)$	個体(箱)を生産性の大きい順に並べたとき, ランク r までの累積生産性.	3, 32
$\Gamma(x)$	ガンマ関数. $\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt$	12
$B(X, Y)$	ベータ関数. $B(X, Y) = \Gamma(X)\Gamma(Y)/\Gamma(X+Y) = \int_0^1 t^{X-1}(1-t)^{Y-1} dt$	12

2.2 各表現の等価性 [5]

$f(x)$ はその定義により,

$$f(x) \propto -dr/dx$$

であるから, この関係を式(2)に適用して, 式(1)と(2)が等価であることが導かれる. $\rho=1+\sigma^{-1}$ となるので, $\sigma=1$ なら $\rho=2$ となる.

また, $x(r)$ は $S(r)$ の微分形であることも定義上明らかであるから, 式(2)と(3)を比較して, (3)は(2)において $\sigma=1$ の場合に相当することがわかる.

3. 確率モデル

3.1 玉入れのモデル

1回1回の試行を, 多数の箱に向かって次々とボールを投げ込んでゆく過程と見なし(ボールは必ずどれかの箱に入る), 全部で N 個のボールを投げ込んだ時, x 個のボールが入った箱(生起回数 x の個体)の存在確率を $f_N(x)$ とする.

さて, これに $(N+1)$ 番目のボールを投げ込んだ時, そのボールがある特定の箱に入る確率が,

その箱にすでに入っているボールの数 x に依存するとしよう. この確率を $p_N(x)$ とすると, 次のような漸化方程式が成り立つ. ただし M は箱の総数である.

$$(4) \quad M \cdot f_{N+1}(x) - M \cdot f_N(x) = p_N(x-1) \times \{M \cdot f_N(x-1)\} - p_N(x) \{M \cdot f_N(x)\} \quad (x \geq 1)$$

$$(4') \quad M \cdot f_{N+1}(x) - M \cdot f_N(x) = -p_N(x) \times \{M \cdot f_N(x)\} \quad (x=0)$$

左辺の N の差分は微分に置き換えてもほとんど問題がないので,

$$(5) \quad \partial f_N(x)/\partial N = p_N(x-1)f_N(x-1) - p_N(x)f_N(x) \quad (x \geq 1)$$

$$(5') \quad \partial f_N(x)/\partial N = -p_N(x)f_N(x) \quad (x=0)$$

$p_N(x)$ の形を与えれば, この式から確率分布関数 $f_N(x)$ を求めることが原理的には可能である. したがって, この式および $p_N(x)$ の形が, 確率分布を導く鍵となる. この他に, 確率の性質から当然次の式が成り立つ.

$$(6) \quad \sum_{x=0}^{\infty} f_N(x) = 1$$

$$(7) \sum_{x=0}^{\infty} x \cdot f_N(x) = N/M$$

$$(8) \sum_{x=0}^{\infty} p_N(x) \cdot f_N(x) = 1/M$$

3.2 Poisson 分布の導出

ボールが飛び込む確率が、箱の中のボールの数 x によらない (すなわち、各試行はそれ以前の試行に独立) とすれば、 $p_N(x)$ は定数であり、式(8)と(6)から、

$$(9) p_N(x) = 1/M$$

となる。これを式(5)、(5')に代入した時の解は、

$$(10) f_N(x) = \exp(-N/M) \cdot (N/M)^x / x! \quad (x \geq 0)$$

となり、Poisson分布になることが確かめられる。

3.3 Zipf-Bradford 分布の導出

次に、ボールがある箱に飛び込む確率が、すでにその箱に入っているボールの数 x に比例すると仮定する。すなわち、

$$(11) p_N(x) = K_N(x/N)$$

経済学者の Simon は、この仮定に立って漸化方程式(5)から確率分布関数 $f_N(x)$ を導いた[3,8]。それは、次のようなベータ関数の形をしており、Yule 分布とも呼ばれている。

$$(12) f_N(x) \propto B(x, \rho) = \Gamma(\rho) \Gamma(x) / \Gamma(x + \rho)$$

ここに $B(x, \rho)$ はベータ関数、 $\Gamma(x)$ 等はガンマ関数であり、 ρ はある制約条件により決まる定数 ($\rho > 1$) である。 x が大きい領域では式(12)は、

$$(13) f_N(x) \propto x^{-\rho} \quad (\rho > 1)$$

となり、Zipf-Bradford分布の式(1)と一致する。

なお、最近 Eto と Makino は、企業が保有する特許の数の分布に対し、多少違ったモデルから(12)と同じ Yule 分布関数を導いている [2]。

3.4 「偶然的」な過程と「人為的」な過程に対する確率分布

ある箱の中に 1 個のボールが飛び込む確率 $p_N(x)$ が、 x によらず一定とした時、Poisson 分布が導かれ、 x に比例とした時 Zipf-Bradford 分布が導かれることがわかった。後者の場合は、すでに多数のボールを収容している個体ほど、新しく追加されるボールを高い可能性で受け入れる

という現象を、最も単純にモデル化した結果である。

Poisson 分布が「偶然的」な過程から生ずるのに対して、このような現象は何らかの「人為的」な過程がかかわっていることに由来すると考えられる。たとえば、論文(ボール)が多くの雑誌(箱)にまったく偶然に投稿されたら、また人間が国土中にまったくランダムに散在していたら、雑誌の掲載論文の分布も、都市の人口の分布も Poisson 分布になるはずである。しかし、より著名な(投稿数の多い)雑誌にますます投稿がなされ、より大都市を目がけて人が集まる結果が、Zipf-Bradford 型の分布になることを、上記の単純なモデルは示している。

Poisson 分布と Zipf-Bradford 分布を導くモデルを図 1 に図解的に示した。

4. 確率モデルの改良

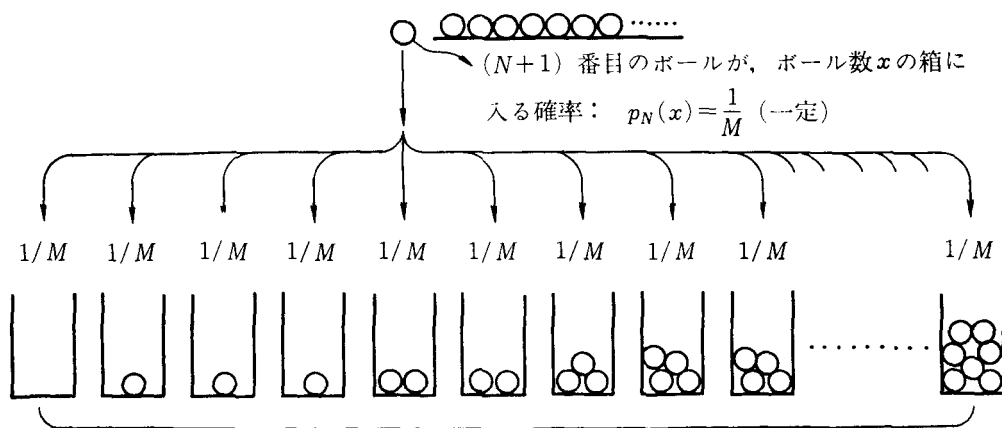
4.1 モデルの問題点

3. で説明したモデルにより、新たなボールが特定の箱に入る確率 $p_N(x)$ が x に比例するとした時、Zipf-Bradford 型の分布が導かれることがわかったが、このモデルには次のような難点がある。

- ① 空の箱 ($x=0$) に新たなボールが入る確率は式(11)より 0 なので、これらの箱は永久に空でありつづけることになる。この不合理を避けるため、Simon [8] はまず空の箱か空でない箱にボールが飛び込む確率比を $a : 1-a$ とし、空でない個々の箱に入る確率を式(11)にしたがうとした (式(12)の ρ はこの a と一意的関係がある)。しかし、この仮定はきわめて人為的であり、空でない箱の総数 M_1 が、ボールの総数 N に比例して増加するという非現実的な面が生ずる。(たとえば、ある小説の最初の 10 ページに現われる語彙の総数に比べて、次の 10 ページに現われる新しい語彙数は少なく、さらに次の 10 ページではもっと減

Poisson 分布を導くモデル

偶然的な現象——1回1回の試行が独立で、他の試行結果に左右されない。



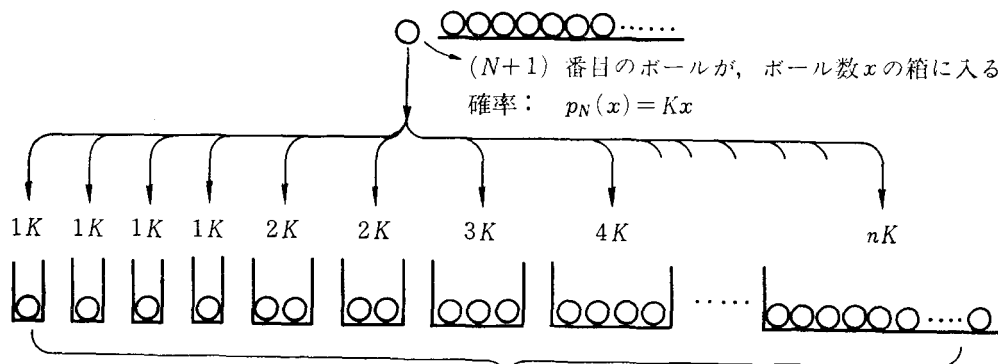
M 個の箱に入った合計 N 個のボール

モデルの解: x 個のボールが入っている箱の数の確率: $f_N(x)$

$$f_N(x) = \exp\left(-\frac{N}{M}\right) \cdot \left(\frac{N}{M}\right)^x / x !$$

Zipf 分布を導くモデル

人為的な現象—— $(N+1)$ 回目の試行における生起確率は、それまでの N 回の結果 (度数分布) に左右される。



M 個の箱に入った合計 N 個のボール

モデルの解: x 個のボールが入っている箱の数の確率: $f_N(x)$

$$f_N(x) = k_N \cdot \frac{x-1}{x-1+\rho} \cdot \frac{x-2}{x-2+\rho} \cdots \frac{1}{1+\rho} = k_N \frac{\Gamma(1+\rho) \cdot \Gamma(x)}{\Gamma(x+\rho)}$$

$$x \text{ が十分大きければ, } f_N(x) \simeq k_N \Gamma(1+\rho) \cdot x^{-\rho}$$

図 1 Poisson 分布と Zipf-Bradford 分布のモデル

るだろう。)筆者はこの仮定をいくぶん緩めてもう少し現実的なものとしたが、やはり十分ではなく、 $f_N(x)$ は $x \rightarrow 0$ で発散するという(12)の形は変わらなかった。

- ② (12)を得る過程で、 $f_N(x)$ は N にのみ依存する部分と x のみに依存する部分の積で表わされるという仮定を置いているが、これも自明ではない。

このモデルでも、想定し得る箱(個体)の総数が事実上無限である場合は問題は少なく、観測される多くのデータをよく説明する。冒頭に挙げた諸例において、箱の総数 M に相当する、あらゆる語彙の数、世界中の雑誌の数や研究者の数などは1つの文章サンプル中の語彙数や、ある主題のサンプル文献集合に含まれる雑誌数や著者数に含まれれば無限に多いであろうから、上記のモデルから導かれる Zipf-Bradford 分布(12)あるいは(13)がよく成立するものと考えられる。また、都市の人口分布や、世帯や企業の収入分布では、 $x=0$ の近くの分布を問題にすることはないから、これも Zipf-Bradford 分布で説明できる。しかし次のような場合は M は事実上も有限であり、 $x=0$ の個体に対しても有効な確率分布関数が必要とされよう。

- a) ある図書館に所蔵する雑誌(その総数が M)に対する一定期間内の利用(複写, 貸出等)の回数の分布
b) 文献データベース中で、統制キーワードや分類コード(その語彙数が M)が索引されている記事数の分布

4.2 偶然性と人為性を統合した改良確率モデル [6]

3.1で導入したモデルに対して、 $(N+1)$ 個目のボールがある特定の箱に入る確率 $p_N(x)$ が、 x に線形に依存するとする。すなわち、

$$(14) \quad p_N(x) = K\{(x/N) + \alpha\}$$

容易にわかるように、この式は(9)と(11)の折衷であり、第1項が「人為的」な因子、第2項が

「偶然的」な因子を表わしている。

式(14)を式(8)に代入し、(6)と(7)を使うと、定数 K が次のように定まる。

$$(15) \quad K = 1/(1 + \alpha M)$$

この(15)および(14)を式(5)および(5')に代入すれば、漸化方程式として次式が得られる。

$$(16) \quad \frac{\partial f_N(x)}{\partial N} = \frac{1}{1 + \alpha M} \left\{ \left(\frac{x-1}{N} + \alpha \right) f_N(x-1) - \left(\frac{x}{N} + \alpha \right) f_N(x) \right\} \quad (x \geq 1)$$

$$(16') \quad \frac{\partial f_N(0)}{\partial N} = -\frac{\alpha}{1 + \alpha M} f_N(0)$$

ここで、 α , N に変えて次のパラメータ

$$(17) \quad \theta = 1 + \alpha M$$

$$(18) \quad \nu = \alpha N$$

を導入し、これにともなって $f_N(x)$ も $f_\nu(x)$ と記述することとすれば、式(16), (16')は次のようになる。

$$(19) \quad \frac{\partial f_\nu(x)}{\partial \nu} = \frac{1}{\theta} \left\{ \left(\frac{x-1}{\nu} + 1 \right) f_\nu(x-1) - \left(\frac{x}{\nu} + 1 \right) f_\nu(x) \right\} \quad (x \geq 1)$$

$$(19') \quad \frac{\partial f_\nu(0)}{\partial \nu} = -\frac{f_\nu(0)}{\theta}$$

この漸化方程式の解 $f_\nu(x)$ が求める確率分布関数である。

4.3 改良モデルに対する確率分布関数

4.3.1 一般解

まず式(19')から $f_\nu(0)$ を求める。 $\nu=0$ (すなわち $N=0$)ではすべての箱が空であるから $f_\nu(0)=1$ という初期条件を用いて、

$$(20) \quad f_\nu(0) = \exp(-\nu/\theta)$$

この解は、 $x=0$ の箱の存在確率は、ボールの総数の増加とともに指数関数的に減少することを示す。

一般の x に対して、式(19)から $f_\nu(x)$ の解析的な解を得ることは、筆者の数学の知識の範囲内では困難である。ご教唆を賜れば幸いである。

そこで、一般解の形を推測するため、式(19)においてまず $x=1$ と置き、 $f_\nu(0)$ に式(20)の解を代

表 2 $x=0 \sim 5$ における $f_\nu(x)$

$$f_\nu(x) = e^{-\nu/\theta} \sum_{i=0}^x A_i(x) \nu^i$$

x	$A_i(x)$					
	$i=0$	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$
0	1	—	—	—	—	—
1	0	$\frac{1}{1+\theta}$	—	—	—	—
2	0	$\frac{1}{(1+\theta)(2+\theta)}$	$\frac{1}{2(1+\theta)^2}$	—	—	—
3	0	$\frac{2}{(1+\theta)(2+\theta)(3+\theta)}$	$\frac{1}{(1+\theta)^2(2+\theta)}$	$\frac{1}{6(1+\theta)^3}$	—	—
4	0	$\frac{6}{(1+\theta)(2+\theta)(3+\theta)(4+\theta)}$	$\frac{11+5\theta}{2(1+\theta)^2(2+\theta)^2(3+\theta)}$	$\frac{1}{2(1+\theta)^3(2+\theta)}$	$\frac{1}{24(1+\theta)^4}$	—
5	0	$\frac{24}{(1+\theta)(2+\theta)(3+\theta)(4+\theta)(5+\theta)}$	$\frac{4(5+2\theta)}{(1+\theta)^2(2+\theta)^2(3+\theta)(4+\theta)}$	$\frac{7+3\theta}{2(1+\theta)^3(2+\theta)^2(3+\theta)}$	$\frac{1}{6(1+\theta)^4(2+\theta)}$	$\frac{1}{120(1+\theta)^5}$

入することによって $f_\nu(1)$ を得、次に式(19)で $x=2$ と置いて $f_\nu(1)$ にいま得た解を得て $f_\nu(2)$ を求める、という過程を逐次的にくりかえして、最初のいくつかの x に対する解を求めた。この結果から $f_\nu(x)$ は一般に次の形になると予想される。

$$(21) \quad f_\nu(x) = \exp(-\nu/\theta) \cdot \sum_{i=1}^x A_i(x) \nu^i \quad (x \geq 1)$$

係数 $A_i(x)$ の $i=1, 2, x, x-1, x-2$ の場合の式を以下に示す。($\Gamma(x)$, $B(x, y)$ はそれぞれガンマ関数、ベータ関数である)

$$(22) \quad A_1(x) = B(x, 1+\theta) = \Gamma(1+\theta)\Gamma(x)/\Gamma(x+1+\theta)$$

$$(23) \quad A_2(x) = B(\theta, 1+\theta) \cdot B(x, 1+2\theta) \cdot \sum_{j=1}^{x-1} \{j \cdot B(\theta, j+1+\theta)\}^{-1}$$

$$(24) \quad A_x(x) = \{x! (1+\theta)^x\}^{-1}$$

$$(25) \quad A_{x-1}(x) = \{(x-2)! (1+\theta)^{x-1}(2+\theta)\}^{-1}$$

$$(26) \quad A_{x-2}(x) = \frac{(3+\theta)x + (\theta-1)}{2(x-3)! (1+\theta)^{x-2}(2+\theta)^2(3+\theta)}$$

式(20)~(26)にもとづき、 $x=0 \sim 5$ における $f_\nu(x)$ の解を表 2 に示す。

4.3.2 $\nu \gg 1$ における漸近解

式(21)の右辺において ν^x と ν^{x-1} の項のみ残し、 $A_x(x)$, $A_{x-1}(x)$ に式(24), (25)を代入すると、次の漸近解が得られる。

$$(27) \quad f_{\nu \gg 1}(x) \simeq P_\mu(x) \times$$

$$\left\{ 1 + \frac{x(x-1)}{\mu(2+\theta)} \right\} \cdot \exp(-\mu/\theta) \quad (x \geq 0)$$

ここに $P_\mu(x)$ は Poisson 分布関数

$$P_\mu(x) = e^{-\mu} \mu^x / x!$$

であり、パラメータ μ は、

$$(28) \quad \mu = \nu / (1+\theta)$$

である。すなわち、 $\nu \gg 1$ での $f_\nu(x)$ の式(27)は、Poisson 分布関数に補正因子を乗じた形をしており、 $\nu \rightarrow \infty$ の極限において Poisson 分布と一致する。 $\nu \gg 1$ の条件は、式(18)と(14)より、1回の試行の確率関数 $p_N(x)$ の第2項(偶然因子)が支配的である場合に相当することがわかる。

4.3.3 $\nu \ll 1$ における漸近解

式(21)の $i=1$ の項のみをとり、 $A_1(x)$ に式(22)を代入すると、 $x \geq 1$ に対する漸近解が得られる。 $x=0$ に対する解(20)と合わせて、

$$(29) \quad f_{\nu \ll 1}(x) = \begin{cases} \exp(-\nu/\theta) & (x=0) \\ \nu \exp(-\nu/\theta) B(x, 1+\theta) \\ \simeq \nu \exp(-\nu/\theta) \Gamma(1+\theta) / x^{1+\theta} & (x \geq 1) \end{cases}$$

$\nu \ll 1$ の条件は、式(14)の第1項、すなわち人為因子が支配的な場合に相当する。この漸近解は、Simon が得た Yule 分布(12)あるいはその近似式である Zipf-Bradford 分布(13)とほぼ一致しているが、次の点で優れている。

① $x=0$ で発散しない確率分布である。

② x の指数 $\rho(=1+\theta)$ に物理的意味を与えた。式(17)から、

$$(30) \quad \rho = 1 + \theta = 2 + \alpha M$$

で、 ρ は偶然因子 α と箱の総数 M に関係づけられる。ここで与えられている制約条件 $\nu = \alpha N \ll 1$ からみて、 αM もそれほど大きくないと考えられるので、指数 ρ はほぼ 2 か、それよりやや大きい値ということになる。しかし経験的に得られる ρ は 1.5~2 程度の例が多く、理論とのあいだにやや不一致が見られる。

4.4 $\nu \ll 1$ における分布の代替的表現

式(29)は、確率モデルの $\nu \ll 1$ の極限として、Zipf-Bradford 分布(1)を理論的に導いたものだが、最初に述べたように、この分布に相当するランカー生起回数型表現 $x(r)$ や、ランカー累積生起回数型表現 $S(r)$ も情報科学ではよく用いられる(式(2)および(3)に相当)。定義により、

$$(31) \quad r = \sum_{x=x(r)}^{x_m} M \cdot f_v(x)$$

$$(32) \quad S(r) = \sum_{x=x(r)}^{x_m} x M \cdot f_v(x)$$

であるから、これらの式の $f_v(x)$ に(29)を代入することにより $x(r)$, $S(r)$ を得ることができる。

(式(31), (32)の x_m は、最高ランク、すなわち $r=1$ の箱に入っているボールの数でボールの総数 N により決まる。) 詳細は文献 [6] にゆずり、結果のみを記す。

$$(33) \quad x(r) \simeq [M\nu e^{-\nu/\theta} \Gamma(\theta)]^{1/\theta} / r^{1/\theta}$$

$$(34) \quad S(r) \simeq \begin{cases} [\theta/(\theta-1)] [M\nu e^{-\nu/\theta} \Gamma(\theta)]^{1/\theta} \cdot r^{(\theta-1)/\theta} & (\theta > 1) \\ M\nu e^{-\nu} \ln\{1 + [(x_m+1)/M\nu e^{-\nu}]\} & (\theta = 1) \end{cases}$$

式(33)は式(2)に対応し、式(34)の $\theta=1$ の場合が Bradford 分布関数(3)に一致する。(33)と(34)は r があまり大きくない(ランクの高い)領域での近似式である。実測のデータに対して $S(r)$ を $\ln r$ に対してプロットした時、図2に示すように、J字型になる場合とS字型になる場合が報告されているが、 r の大きいところでの挙動を考察することにより、パラメータ θ のある臨界値を境にし

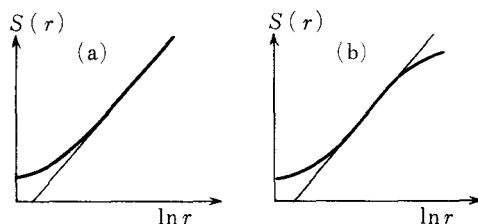


図2 $\ln r - S(r)$ プロットで経験的に得られる2つの形
(a) J字型 (b) S字型

て、この相異が現われることも証明される。

5. おわりに

本稿に述べたような理論的解析が、情報科学においてどのような応用領域をもつか、1, 2の例で紹介したい。

Bradfordが英国のScience Museum Libraryの館長であったことでもわかるように Zipf-Bradford 分布は、特定分野の雑誌の収集計画の立案に参考となる。 $S(r)$ の形から、その分野の中核誌、準中核誌、周辺誌の数を予測し、たとえば、利用の80%とか90%を満たすには何種類の雑誌を収集すべきかの分析から、コスト効果を考えて収集誌を決めることになる。

文献データベースの検索のためには、その中の索引語、標題、抄録等の中の用語を抽出したキーワードファイルを作成するのが普通である。検索システムの必要記憶装置や処理能力を設けるうえで、たとえば100万件あるいは1000万件の文献を蓄積した時、キーワードの語彙数とそれらの度数分布を予想することは重要である。ここで述べたモデルは、比較的少量の文献集合における分布の形から、大量の集合のそれを予測するのに役立つ。

Zipf-Bradford分布は、Poisson分布や二項分布に比較して、特定の個体に生起度数が集中し、そのために分散の大きい分布になる。このため、実用の世界でこの種の分布を取り扱うシステムには、コスト効果の面で厄介な問題が生ずる。情報の取り扱いに計量情報学が必要とされる所以であ

る。

参 考 文 献

- [1] Bradford, S. C.: *Documentation*, Crosby Lockwood, London, 1948
- [2] Eto, H. and Makino, K.: Stochastic Model for Innovation and Resulting Skew Distribution for Technological Concentration with Verification in Japanese Industry, *Scientometrics*, Vol. 5, No. 4 (1983), 219-243
- [3] Ijiri, Y. and Simon, H. A.: *Skew Distributions and the Sizes of Business Firms*, North-Holland, Amsterdam, 1977
- [4] Lotka, A. J.: The Frequency Distribution of Scientific Productivity. *J. Washington Acad. Sci.*, Vol. 16, No. 12 (1926), 317-323
- [5] 小野寺夏生: "Bibliostatistics" - 情報現象の統計学的説明, 情報管理, Vol. 21, No. 10 (1979) 782-802
- [6] 小野寺夏生: Zipf 分布と Poisson 分布を統合する確率モデル, 第20回情報科学技術研究会発表論文集, 日本科学技術情報センター, 東京, 1984, pp. 173-183
- [7] Pareto, V.: *Cours d'économie politique*, Vol. 2, Sec. 3, Lausanne, 1897
- [8] Simon, H. A.: On a Class of Skew Distribution Functions. *Biometrika*, Vol. 42 (1955), 425-440
- [9] Willis, J. C.: *Age and Area, a Study in Geographical Distribution and Origin of Species*. Cambridge University Press, 1922
- [10] Zipf, G. K.: *Human Behavior and the Principle of Least Effort*, Hafner Publishing Co., New York, 1972

新時代のコンピュータ総合誌

Computer Today

● 3月号/好評発売中

定価880円

コンピュータと自然言語

コンピュータは日常言語をどこまで理解できるか

自然言語処理の目的と方法	辻井潤一
ロボットと自然言語	森戸 誠
日本語処理の問題点	佐竹秀雄・中野 洋
シンタクスとセマンティクス	今井邦彦
機械翻訳の実際	牧野武則
音声認識: 話し言葉をどう処理するか	小畑秀文
パーソナル・コンピュータによる自然言語処理	草薙 裕
状況意味論	白井英俊
GB 理論	原口庄輔

お得な年間購読のおすすめ 年間5000円(6冊)

● 既刊 1月号

エキスパートシステムズ

エキスパートシステムの構造と課題を明らかにし, 人工知能理論の役割, アーキテクチャ, 言語およびツールを解説し, 実例を紹介する。

数 理 科 学

4月号予告

定価 880 円

特集 = 逆 問 題

科学において逆問題は成立するか

吉田夏彦・林 一・坂本百大
今井 功・香原志勢

人間原理	佐藤 文隆
日本人の起源を考える	塩原 和郎
微分方程式の逆問題	鈴木 貴
ソリトンと逆散乱法	加藤 祐輔
宇宙X線の観測と逆問題	牧島 一夫
超解像における逆問題とその解法	安藤 繁
医学における逆問題	田中 博
資源探査における逆問題	岡部 政之
地下水流の解析と逆問題	田中 豊・中屋真司

数理科学・別冊

形・フラクタル 4月上旬
発売予定

ソリトン 好評発売中
定価2000円

サイエンス社

東京都千代田区神田須田町2-4 安部徳ビル

☎03(256)1091 振替 東京7-2387