情 報 学 基 礎 21－3
（１９９１．５．２８）

大規模データベースの構成と利用に関する統計的特性

小野寺 夏生
日本科学技術情報センター

大規模なデータベースの編成及び利用の場面で得られるいくつかの大きな統計サンプルにより、「ゆがんだ」度数分布の実例を示した。一般的に、ある生産要素の生産性の増加確率が、それまでにその要素が貯えている生産性に線形に依存するという非独立事象の確率モデルから、このような分布の説明を試みた。現実の分布のパターンは、自然語キーワードの出現頻度分布のように生産要素数が無限と想定される場合と、統制索引語の付与頻度の分布のように全生産要素数が固定される場合に大別できる。モデルから導かれる分布関数は、前者の場合ベータ関数型の分布、後者の場合負の二項分布になり、実際に観測される分布をよく説明する。

STATISTICAL NATURE OBSERVED IN THE STRUCTURE AND THE USE OF LARGE-SCALE DATABASES

Natsuo Onodera

The Japan Information Center of Science and Technology

5-2, Nagatacho 2 chome, Chioyoda-ku, Tokyo 100, Japan

Some examples of the 'skew' distribution observed in large-scale database activities are given. Such distribution is attempted to be elucidated generally by a stochastic model for not-independent events, where the probability of increase in the productivity for a particular element depends linearly on the productivity the element holds in that moment. Actual frequency distribution patterns seem to be divided into the cases of infinite and fixed numbers of elements. The distribution function derived from the model is a beta-function type or a negative binomial type for the infinite or fixed elements case, respectively. These explain sufficiently the observed data sets.

1．情報現象に見られる「ゆがんだ(skew)」度数分布

情報の発生や利用に関する何らかの「生産性(productivity)」の指標xには、次のような「ゆがんだ(skew)」度数分布を示すものが多い。

① 非対称性が非常に大きく、xの小さい側にほとんどの度数が集中、

② 一方、xの大きいところでの度数もなかなか0に収れんせず、長い尾を引く。

この種の分布で昔からよく知られているものに以下のような例がある。

(a) 特定のテキスト中でx回使われている語の種数n(x) (Zipf, 1949)

(b) ある期間内に特定のテーマの論文をx件掲載した雑誌数n(x) (Bradford, 1948)

(c) ある期間内にx件の論文を発表した著者の数n(x) (Lotka, 1926)

但し、このうち生産性xを持つ生産要素の度数n(x)の形で分布を表したのはLotkaだけである。彼は上記の分布に対し、

$$n(x) \propto x^{-p}, \quad p \sim 2 \qquad (1)$$

を提案した。他の2例では、それぞれ次のような表現により分布が示されている。

(a) Zipfによれば、テキスト中の語をその使用頻度順に並べた時、上位からr番目の語の使用頻度x(r)は次式に従う（ランク－生産性関係）。

$$x(r) \propto r^{-q}, \quad q \sim 1 \qquad (2)$$

(b) Bradfordによれば、雑誌をその生産性（あるテーマの論文の掲載数）の順に並べた時、上位r番目までの雑誌の累積生産性S(r)は次式に従う（ランク－累積生産性関係）。

$$S(r) \propto \log(1+\beta r) \qquad (3)$$

その後の多数の事例報告（上記の3つの現象以外にも、図書館における蔵書の閲覧や貸出の回数、データベースにおけるキーワードの使用回数、雑誌の引用回数等の分布への適用例がある）においても、これら3種の表現がそれぞれに使用されているが、以下の関係を用いれば、この3つの式がほぼ等価関係にあることが容易に証明される。

$$S(r) = \Sigma_{r'=1}^{r} x(r')$$
$$r = \Sigma_{x=x(r)}^{x_m} n(x)$$
$$q = 1/(p-1) \qquad (4)$$

（但し(3)は(2)でq=1とした時得られる。q≠1だとS(r)は別の形になる）

ZipfとBradfordの式は度数の大きな生産要素に注目する調査に、Lotkaの式はむしろ小さい度数の生産要素に重点を置く調査に利用されることが多い。いずれにせよ、多くの情報現象がこのような収れん性の悪い分布を示すことは、度数の大きい要素だけ、あるいは小さい要素だけを考慮する訳に行かないことを意味し、図書館サービスやデータベースサービスの設計や実施を困難あるいは非効率にする本質的原因となっている。

このような事情から、「ゆがんだ」分布がなぜ生ずるかを理解し、その理解の上に立って現実の分布を定量的に解釈・推定・予測できることが望ましい。これについてはこれまでにいろいろの試みがなされているが、実際に得られるサンプルが小規模であったり、ある特定の現象のみを対象にしている場合が多い。ここでは、ゆがんだ分布を導く一般的確率モデルを提示し、大規模データベースの構成と利用に関する大きなサンプルにそれを適用して定量的信頼性を検証する試みについて述べることとする。

2．大規模データベースに関わる分布の例

2．1 2通りのゆがんだ分布

ゆがんだ分布を次の2種に大別することが、

その整理と理解を容易にすると思われる。

(1) 潜在的な生産要素は無限母集団をなすと想定され、その中の0でない生産性をもつ要素だけが実際の観測にかかると見なせる場合。冒頭に挙げたZipf, Bradford, Lotkaの分布は、それぞれの母集団である語彙の総集合、あるいは関連の雑誌や研究者の集合は無限と考えられ、この場合に相当する。

(2) 対象の全生産要素が固定されている場合。統制キーワードの付与回数の分布や、図書館における蔵書の利用回数の分布では、キーワードの全語彙や蔵書の総数は固定されており、この場合に相当する。

2．2　生産要素数無限の場合の分布の例

(1) データベース中のタイトル語の出現度数

JICST科学技術文献ファイル（以下JICSTファイルという）は、JOISでのオンラインサービスのために文献タイトルから自然語キーワードを自動抽出している。1981年4月から1983年11月までの2年8ヶ月間のファイル（総記事数1,142,816件）から抽出されたタイトル語の頻度（出現した記事数）分布を**図1**（Lotka型）と**図2**（Zipf型）に示す。中間のランクのデータが欠けているが、この分布は式(1)及び(2)にきわめてよく従い、それぞれの指数パラメータはほぼp=2, q=1であることが解る。

(2) 米国におけるデータベースの利用度

IMI Reportは、米国内におけるテキスト系の主要14オンラインサービスの推定利用統計を四半期毎に公表している。1988年の最終四半期の統計から、各データベースの利用時間分布を求めた（**図3**及び**図4**）。これらの図から、分布が式(1)及び(2)によく従うことが解る。しかし、回帰解析により得られる各々の指数パラメータはp=1.562, q=1.365となり、pとqの間の理論的関係(4)とはやや合わない。

2．3　生産要素数有限の場合の分布の例

(1) データベース中の主題分類の付与頻度

文献データベースに付与される主題分類は一種の統制語索引であり、その付与頻度の分布は、全生産要素数（即ち主題分類コードの個数）が有限の典型例と考えられる。

JICSTファイルに用いられている主題分類（JICST科学技術分類）コードの付与頻度xの分布を**図5**に示す。この図は、1986年8月に更新された48,268記事における分布である。

自然語の場合（**2．2**参照）のように両対数プロットが直線状にならず、上に凸の湾曲した形になる。

(2) 所蔵雑誌に対する複写要求回数

図6は、1986年度のJICSTファイルへの採択雑誌に対して1987年4月から1988年3月までの1年間に要求された複写の回数を、雑誌別に集計した結果である。この場合も採択誌の総数は固定されているから、有限の生産要素の例として扱える。**図5**と同様に、両対数目盛で湾曲した分布形になる。

3．ゆがんだ分布を導くモデル

3．1　非独立事象に対する確率モデル

2．で示したようなゆがんだ分布が生ずる過程を理解するために、多数の「箱」（各生産要素を表す）に次々と「ボール」を投げ込んで行く確率モデルを考える。1回の試行（1個のボールの投げ込み）でボールはどれか1つの箱に入る。個々の試行がそれ以前の試行の結果に独立で、M_0個の箱のどれにも同じ確率でボールが入るとすれば、N個のボールが投げ込まれた時点でx個のボールが入っている（即ち生産性がxの）箱の数$n_N(x)$は、以下のようなポアッソン分布に従うであろう（但しM_0, Nはともに十分大きな数とする）。

$$n_N(x)=M_0\cdot{}_NC_x(1/M_0)^x(1-(1/M_0))^{N-x}$$
$$=M_0\exp(-\mu)\mu^x/x! \qquad (5)$$
但し$\mu=N/M_0$

この分布も非対称ではあるが、xの大きいところで極めて速やかに減衰するので、ここで問題とするゆがんだ分布ではない。

しかし、独立事象の仮定を捨て、「それま

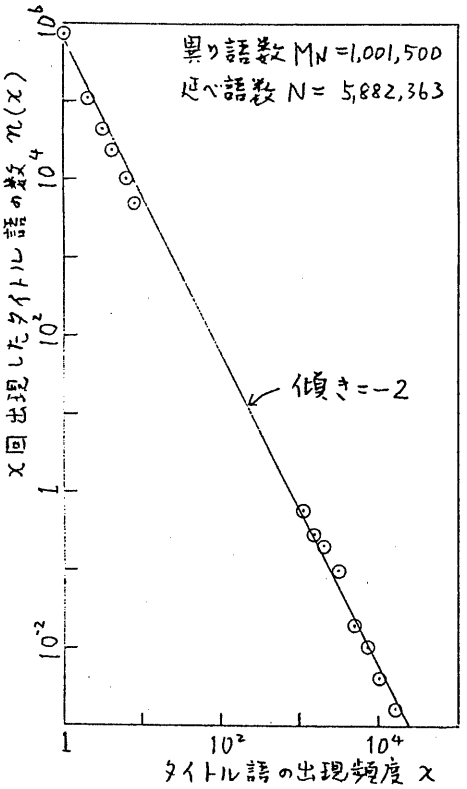

図1　タイトル語の出現頻度分布（JICSTファイル 1981.4-1983.11）

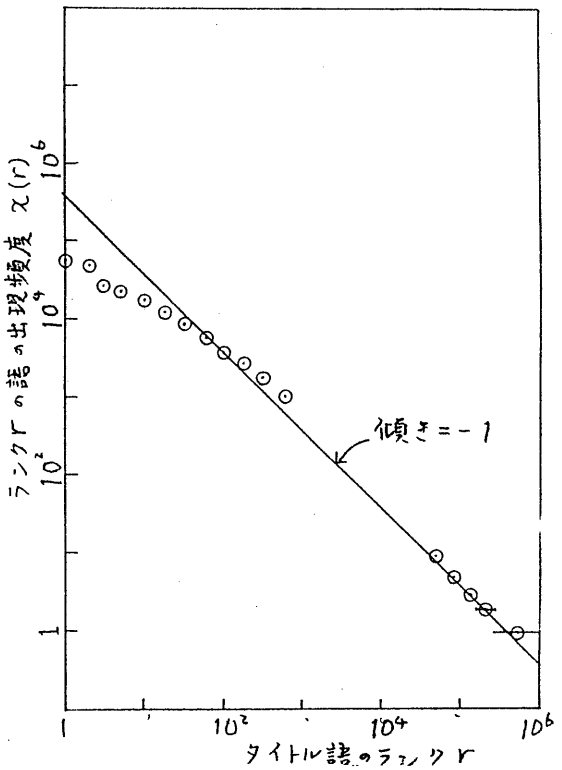

図2　タイトル語のランク−出現頻度関係（同左）

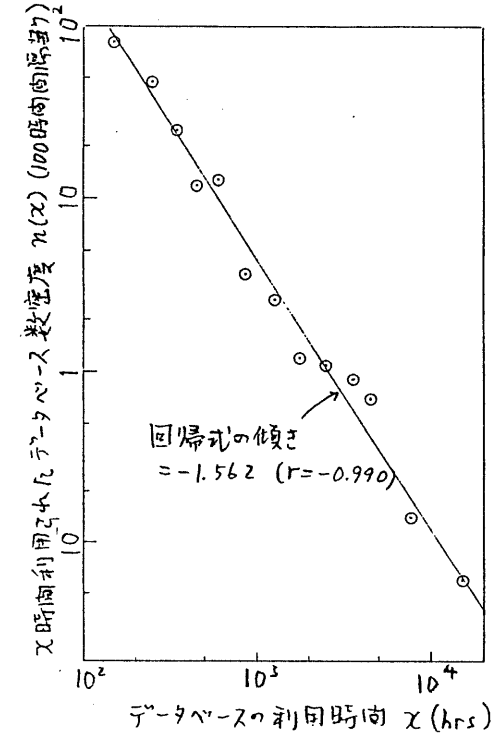

図3　米国におけるオンラインデータベースの利用時間の分布

(1988.10-12, IMI Repotによる)

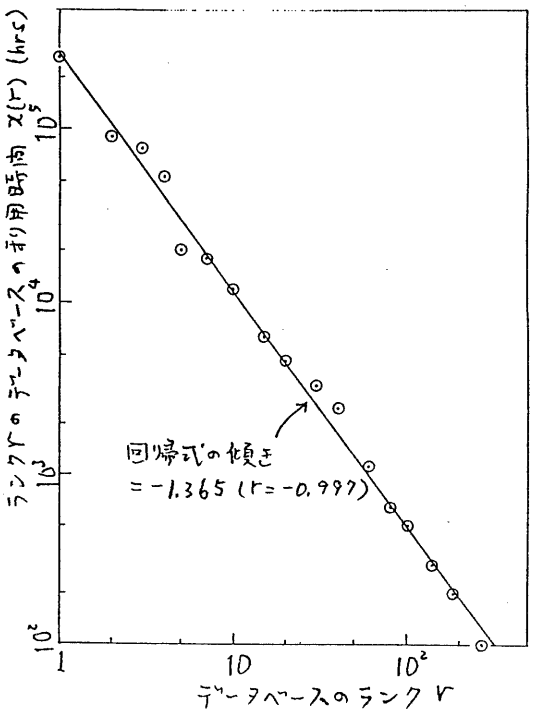

図4　米国におけるオンラインデータベースのランク−利用時間関係（同左）

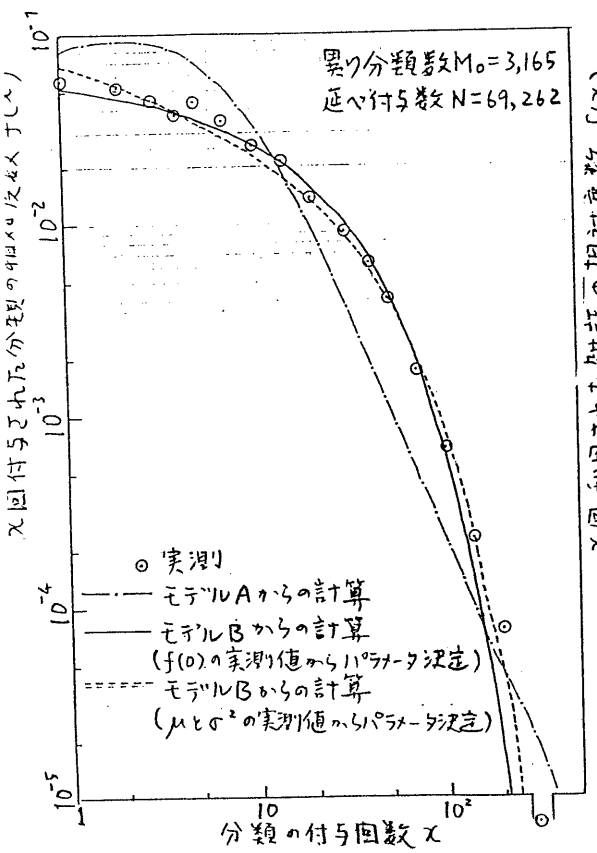

図5　主題分類コードの付与回数の分布

(サンプル C1：JICSTファイル 1986.8)

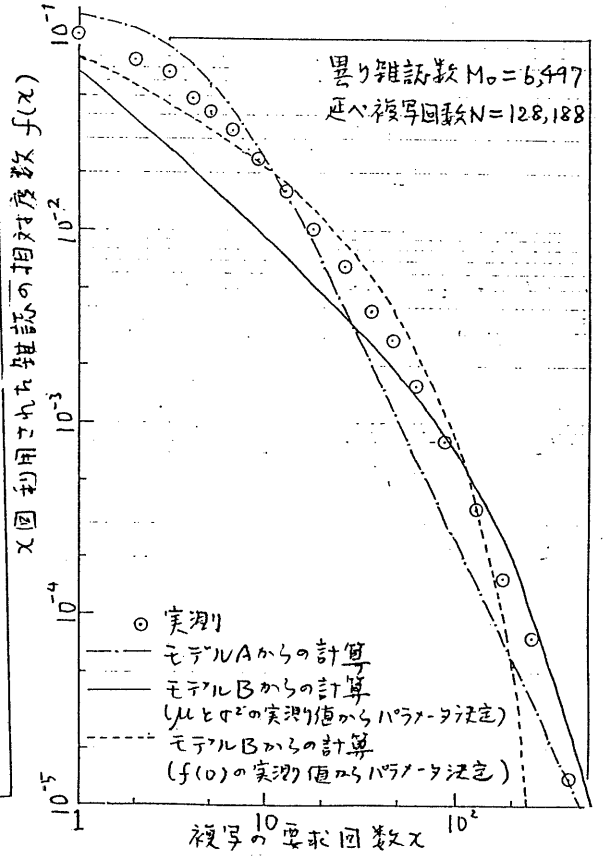

図6　雑誌に対する複写要求回数の分布

(対象：1986年度JICSTファイル採択誌、期間：1987.4.1-1988.3.31)

でに既に多数のボールが入っている箱には次の試行でボールが入る確率が高い」と仮定すると分布はどうなるであろうか。具体的にいうとこの仮定は、始終使われている語の使用確率は高く、よく投稿（あるいは利用）されている雑誌に対して新しい投稿（利用）も集中し、過去の生産性の高い著者ほど新たな論文を書く可能性が高いということを意味する。情報活動のような偶然より人為が支配する現象においてはこの仮定は妥当（ある意味では当然）であろう。このようなモデルは、「成功が成功を産む」過程とか、「付和雷同型」の挙動とか呼ぶことができる。

上記の仮定を定量化するため、最も単純に、「ある分布状態にある箱の集合に対して新に投げ込まれた1個のボールが個々の箱に入る確率は、その箱に既に入っているボールの数xに比例（または線形依存）する」としてみる。箱の総数が無限と見なせる場合と固定されている場合のそれぞれについて、この仮定からどんな分布が導出されるかを以下に考察する。

3．2　無限個の箱の場合に対するボール数の分布

総数N個のボールが投げ込まれた時点でのボール数xの箱の数を$n_N(x)$とし、(N+1)個目のボールが特定の箱に入る確率を$p_N(x)$とする。このステップにおける$n_N(x)$の変化は、(N+1)番目のボールが$n_N(x)$個の箱のどれかに入ってそれが減る確率と、$n_N(x-1)$個の箱のどれかに入って新たにボール数xの箱になる確率で決まる。この遷移方程式は以下のようになる。

$$\partial n_N(x)/\partial N=p_N(x-1)n_N(x-1)-p_N(x)n_N(x) \quad (6)$$

ここで$p_N(x)$は箱内のボール数xに比例するという仮定

$$p_N(x)=K_N x, \quad x\geqq 1 \quad (7)$$

を適用し、(6)に(7)を代入すると、

$$\partial n_N(x)/\partial N=K_N\{(x-1)n_N(x-1)-xn_N(x)\} \quad (8)$$

但しこの式はx=1に対しては不適である（右辺の第一項が$0\times\infty$の形になる）。x=1の場合は、それまで空であった箱のいずれかにボールが入って新しく観測にかかる箱（その総数をM_Nとする）になる確率がdM_N/dNであることから、

$$\partial n_N(1)/\partial N=(dM_N/dN)-K_N n_N(1) \quad (8')$$

$n_N(x)$の形を決めるためにもう一つ以下のような仮定を置く。箱が無限と見なされる場合は、ある定常状態に達すれば、ボールの総数Nをどんどん増やして行っても$n_N(x)$の分布の形は変化せず、全体的にその形が大きくなるだけとしてよいであろう。このことは$n_N(x)$が

$$n_N(x)=g_N\cdot f(x) \quad (9)$$

と書けることを意味する（分布の形を決めるのはf(x)である）。

(8)に(9)を代入し、f(x)に関する漸化式を解けば、

$$n_N(x)=g_N B(x,1+\rho) \quad (10)$$

となることが示される。ここにB(x, 1+ρ)はベータ関数

$$B(x,1+\rho)=\Gamma(1+\rho)\Gamma(x)/\Gamma(x+1+\rho)$$
$$=(x-1)!/(x+\rho)(x-1+\rho)\cdots(1+\rho)$$

で、xが大きいところでは近似的に

$$B(x,1+\rho)\fallingdotseq\Gamma(1+\rho)/x^{1+\rho} \quad (11)$$

が成立する。ρは値が1前後の正の定数である。(11)はLotkaの分布関数(1)と同じ形をしており、ここにLotka型分布が本節で述べた無限生産要素に対する付和雷同型確率モデルにより説明された。

更に、

$$\Sigma_{x=1}^{x_m} n_N(x)=M_N \quad (12)$$
$$\Sigma_{x=1}^{x_m} x n_N(x)=N \quad (13)$$

の関係及び式(8')を使ってM_N, g_N, K_N, x_mのNに対する依存形を求めることができる。このうち興味のあるM_Nの形については、

$$1/(dM_N/dN)=1+(\mu/\rho) \quad (14)$$
但し$\mu=N/M_N$

であることが導かれる。２．２(1)に例示したタイトル語の度数分布において、データベース蓄積期間の各段階でのNとM_Nのデータが得られているが、このデータについて見ると図7

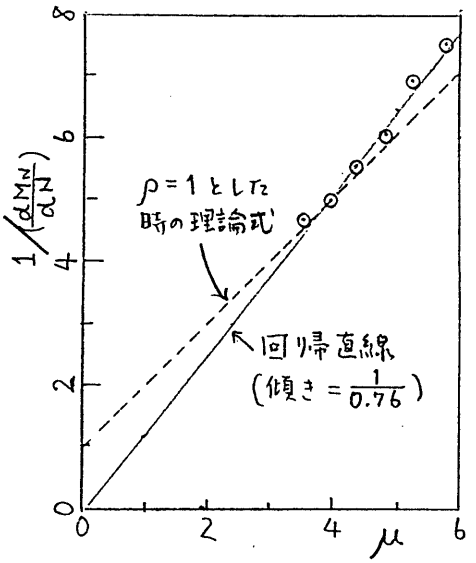

図7　タイトル語の延べ語数Nと異なり語数M_Nの関係
(JICSTファイル 1981.4-1983.11の期間における推移)

のように$1/(dM_N/dN)$とμの間に直線関係が認められ、その傾きから$\rho=0.76$となる（但し図1からは$\rho=p-1=1$となるべきである）。

3．3　有限個の箱の場合におけるボール数の分布

(1) 基本方程式

この場合は、前節の無限個の箱の場合と以下の点で考え方を変える必要がある。

① 観測される箱（総数M_N）がNの増に伴いどんどん湧き出すという考えは不適当である。$x=0$の箱も含めた箱の総数M_0が固定されているという考えに立つ。

② Nの変化に応じて分布関数$n_N(x)$の形そのものが変化する。つまり、$n_N(x)$は(9)のようにNの項とxの項に分離することができない。Nの増加に伴い$n_N(x)$の重心はxの大きい方に移動し、分散も拡がり、場合によっては$n_N(x)$曲線にピークができるだろう。

③ ボールが箱に入る確率はxに比例するとし、$x=0$の箱にボールが入る確率だけは特別扱いとした前節の仮定（(8)及び(8')）はこの場合は不自然である。

このような考えに立つと遷移方程式は以下のようになる。

$$\partial n_N(x)/\partial N=p_N(x-1)n_N(x-1)-p_N(x)n_N(x) \quad (x\geq 1) \quad (15)$$

$$\partial n_N(0)/\partial N=-p_N(0)n_N(0) \quad (x=0) \quad (15')$$

ここで$p_N(x)$に対してxに線形依存（比例ではなく）を仮定し、次の2通りの式を検討する。

モデルA：　$p_N(x)=K_A((x/N)+a)$　　(16-1)

モデルB：　$p_N(x)=K_B(x+b)$　　(16-2)

a, bは、$x=0$の箱にボールが入る確率が0にならないように設定したNによらない定数である。

方程式(15)、(15')に(16-1)または(16-2)を適用し、以下の統計上の関係

$$\Sigma_{x=0}^{xm} n_N(x)=M_0 \quad (17)$$

$$\Sigma_{x=0}^{xm} x n_N(x)=N \quad (18)$$

$$\Sigma_{x=0}^{xm} p_N(x)n_N(x)=1 \quad (19)$$

を使うことによって、モデルA, Bのそれぞれに対し以下の解が導出される。但し、以下では出現度数n_Nの代わりに相対度数

$$f_N(x)=n_N(x)/M_0$$

を求めることとする。

(2) モデルAに対する解

$$f_N(x)=\exp(-\nu/\rho)\Sigma_{i=1}^{x}\nu^{i}A_i(x) \quad (x\geq 1) \quad (20)$$

$$f_N(x)=\exp(-\nu/\rho) \quad (x=0) \quad (20')$$

ここで　$\rho=1+aM_0$　　(21-1)

$\nu=aN$　　(21-2)

$A_i(x)$, $i=1,2,\cdots,x$はある漸化式の解として得られるxの関数である。

式(20)は解析的には表現できない複雑な関数なので、$\nu\ll 1(1/N\gg a)$と$\nu\gg 1(1/N\ll a)$の両極限における近似解を与えることにする。(16-1)を見れば解るように、これらの両極限はそれぞれ、遷移確率$p_N(x)$の比例項が主要な場合と定数項が主要な場合に相当する。

(a) $\nu\ll 1$ の場合の近似解

$$f_N(x)\fallingdotseq \nu\exp(-\nu/\rho)B(x,1+\rho) \quad (22)$$

$B(x,1+\rho)$はベータ関数で、無限個の箱の場合の解(10)と同様な形となる。従ってxが大きいところでは(1)と等価な

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

に近づく。(21-1)から解るように、指数$1+\rho$の値は2より大きい。

(b) $\nu\gg 1$の場合の近似解

この時はポアッソン分布

$$f_N(x)=\exp(-\lambda)\lambda^x/x!$$

に近い形になる($\lambda=\nu/(1+\rho)$)。$p_N(x)$がxにほとんどよらない場合であるから当然と言える。

(3) モデルBに対する解

$$f_N(x)=[\Gamma(x+b)/\Gamma(b)\Gamma(x+1)]\times(b/(\mu+b))^b(\mu/(\mu+b))^x \quad (23)$$

但し $\mu=N/M_0$

これは平均μ、分散$\mu(\mu+b)/b$をもつ負の二項分布である。

(4)実際の分布との比較

2．3に示した2つの例、JICSTファイル中の分類の付与頻度と同ファイル採択誌への複写要求回数の分布に対して、上記のモデルAとBの解(20)と(23)をそれぞれ当てはめた。結果は**図5**、**図6**のグラフ中に示されている。2つのモデルの優劣は**図6**ではそれほど明らかではないが、**図5**ではモデルB（負の二項分布）の方が一致度がよい（どちらのモデルも任意に設定できるパラメータは1個（モデルAのa、モデルBのb）であるからその点の有利不利はない）。

モデルBの優越性は、更に大きいサンプルの分類付与頻度分布に適用してみた時明白になった。**図5**は1ケ月の期間のJICSTファイルに対する結果であるが、4ケ月分及び18ケ月分のファイルに対し、パラメータa,bは1ケ月分のサンプルから得られた値をそのまま使い、総頻度Nだけを実測値に合わせて変化させることにより理論分布を計算した。その結果、モデルAは実測と全く一致しなかったが、モデルBは極めてよく実測を再現した（**図8**）。

以上のことから、箱が有限の場合の確率モデルとして、遷移確率(16-2)に基づいて導かれる負の二項分布が適切であると考えられる。

参考文献

小野寺夏生　情報管理, 21[10], 782-802('79)
Onodera, N. Scientometrics, 14[1-2], 143-159('88)

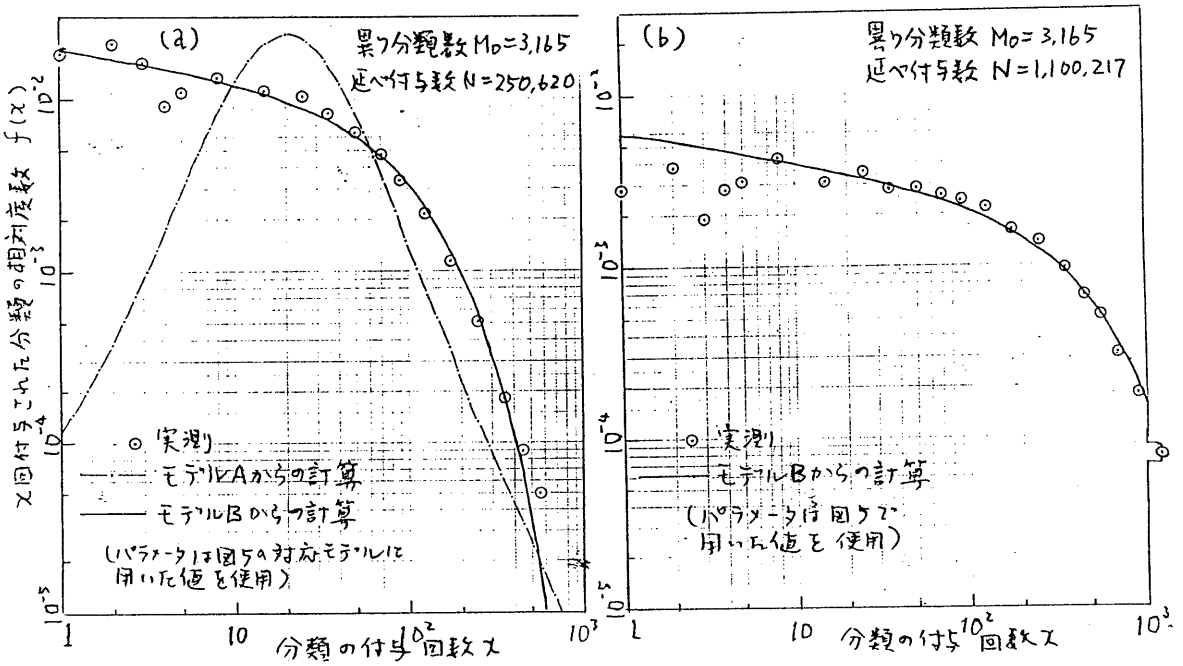

図8　主題分類コードの使用頻度の分布：実測とモデル計算の比較

(a)ｻﾝﾌﾟﾙC4:JICSTﾌｧｲﾙ1986.4-7　　(b)ｻﾝﾌﾟﾙC18:同1985.4-1986.9