数理解析研究所講究録
第1796巻 2012年 26-30

日本人の名前のサイズ頻度分布

Size Frequency Distribution of Japanese Names

早川 良 (Ryo HAYAKAWA), 水口 毅 (Tsuyoshi MIZUGUCHI)
大阪府立大学大学院
工学研究科電子・数物系専攻 数理工学分野
Department of Mathematical Sciences,
Osaka Prefecture University

1 はじめに

日本人には名字と名前がある．名字は苗字，氏，姓などと称されるが，本稿では統一して“姓”と呼ぶことにする．それに対し，名前を“名”と呼ぶことにする．日本には多くの姓があり，その数は10万種以上あると言われている [1]．その中にはありふれたもの (多出姓) も珍しいもの (希少姓) も存在し，その頻度にはばらつきが見られるはずである．では，それらのばらつきの統計的な特徴はどのようなものだろうか．

姓の分布に関する研究は多く行われている．日本人の姓に関する先行研究としてはMiyazima らによるもの，Chida,Mase によるもの，Sato,Seno によるものなどがある [1, 2, 3, 4]．Chida,Mase は 29,727,887 件の姓のデータを解析し，希少姓のサイズと頻度の間にべき則を見出している．また，既存の姓の総数の推定や，Galton-Watson モデルを用いた数値計算で今後の希少姓の減少具合を見積もっている．Miyazima らは人口と総姓数，姓のサイズと頻度，姓のランクとサイズの間にべき則を見出している．

これらに対し，我々は名前の分布に着目した．日本人の姓名が掲載されているデータベースから姓と名のリストを抽出し，その分布を統計的に特徴付け，比較した．その結果，姓と名はそれらのサイズ頻度分布が非常に似通っていることを見出した．姓は誕生時に親のものを継ぐのが原則であるのに対し，名は任意に決められる．また，姓の種類が増えることは稀だが，名は新たなものが日々増え続けている．さらに名は流行の影響を受けるが，姓はそうではない．これらの点で姓と名はその生成プロセスを異にしている．にもかかわらず，分布関数が酷似しているのはなぜだろうか．この疑問に答えるための第一歩として，得られたサイズ頻度分布が形成されるメカニズムを明らかにするため，命名過程をモデル化し，名前を有する仮想的な人口集団を構築し，名前のサイズ頻度分布を測定した．その結果，あるパラメータ領域で実測されたデータと同じ統計的特徴を有するサイズ頻度分布を得ることに成功した [7]．

2 実測

本研究では姓と名のそれぞれに対して実測・解析を行った．使用するデータとして，研究開発支援総合ディレクトリ (ReaD)[5] を用いた．得られたデータは 211,955 人分であった．姓 (名) の区別は漢字表記のみで行なっている．得られたデータには，姓・名に加え

そのカナ表記もあったので,「堀田 (ホッタ)」と「堀田 (ホリタ)」のような同字異音の姓を区別することも可能であるが，本稿では同一の姓として扱っている．逆に同音異字の姓 (例えば「渡辺 (ワタナベ)」と「渡邉 (ワタナベ)」) については，別の姓として扱った．

このデータは外国人の氏名を含むと思われる．本研究では日本人の姓・名の分布に着目するために，姓・名に全角半角のアルファベットおよびカタカナを含むものを消去した．このため漢字を用いた中国人姓や韓国人姓は残っていると考えられる．除いたデータは 12,096 人分であり，最終的に解析したデータは 199,859 人分であった．データの破棄率は 5.71 %であった．なお，総姓数は 19,812，総名数は 33,923 であった．

取得したデータについて，表 1 に示した統計量を計算した．サイズと頻度を両対数でプロットしたものが図 1 である．姓・名ともに，サイズと頻度が希少領域でべき則

$$n(s) \propto s^{-\gamma} \tag{1}$$

が成り立っている．希少領域 ($10^0 < s < 10^2$) でフィッティングをしたところ，べきの指数はそれぞれ $\gamma_{姓} = 1.79$, $\gamma_{名} = 1.84$ であり，分布関数は酷似している．

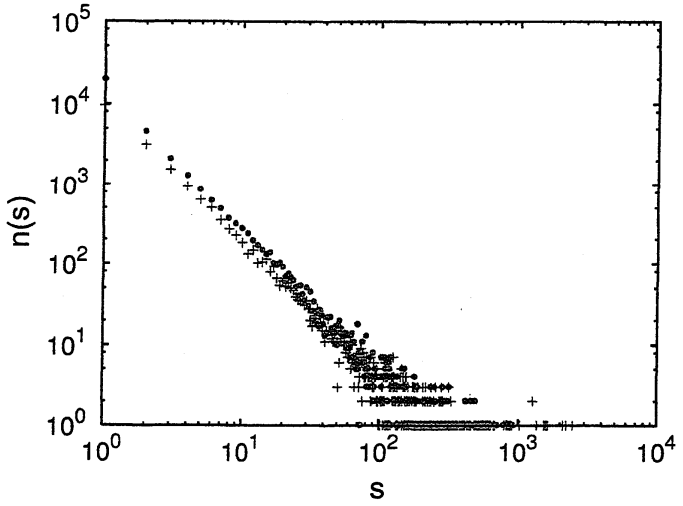

図 1: ReaD の姓・名のサイズと頻度の関係．希少領域においてべき則が見られる．

統計量	記号	説明
総人口	S	データ内の人口
総姓数	N_f	データ内の姓の種類数
総名数	N_g	データ内の名の種類数
サイズ	s_j	姓がjである人の数
頻度	$n(s)$	同じサイズsを持つ姓の種類数

表 1: 本稿で扱う統計量．

3 モデル

前節で述べた通り，姓だけでなく名も，そのサイズ頻度分布が希少領域でべき則を示すということがわかった．継承過程の異なる姓と名が，同様のべきの指数を示したことは非常に興味深いと思われる．姓の継承については Sato,Seno などが Galton-Watson モデルを用いた数値計算を行なっている [1]．そこで本稿では名付けに着目し，いくつかの傾向を取り入れたモデルを立て，べき則を再現することを試みる．

名付けのプロセスは複雑である．これを単純にモデル化することは難しいが，その傾向や制約を考えることは出来る．まず選択傾向として，有名人の名にあやかったり，① 流行の名を付けたり，② 全く新しい名を創作したり，親から一字受け継ぐこと，などがある．逆に制約としては，③ 親と同じ名，兄弟で同じ名は付けないといったものがある．本研究では，③ を排重過程と呼ぶことにする．

これらの傾向をモデル化するために，Yule 過程 [6] を取り入れたモデルを構成しよう．ある個体が誕生した際の名は次のように決まる．ある確率 α で，今までに存在しなかった新名が付く (① と対応)．そうでない場合 (確率 $1-\alpha$) には，既存の名が付く．この時にどの既存名が付くかを，以下のように決めた．s_j をある名 j のサイズとすると，発生した個体に j という既存名が付く確率 $P(j)$ を

$$P(j) = \frac{(s_j)^\beta}{\sum_j (s_j)^\beta} \tag{2}$$

で与える．必ずいずれかの既存名が付くことから

$$\sum_j P(j) = 1 \tag{3}$$

が成立する．ここで β は実数パラメータである．例えば $\beta = 0$ では，すべての名前が同一の確率 $1/N$ で選ばれ，$\beta = 1$ ではいわゆる Yule 過程となる．これに対して $0 < \beta < 1$ では $P(j)$ のサイズ依存度が小さく，また $1 < \beta$ では $P(j)$ のサイズ依存度が大きくなる．このように β は $P(j)$ の既存名サイズへの依存度を左右するパラメータであり，流行の影響を示すパラメータと考えることができる (② と対応)．このため β を流行反映度と呼ぶことにする．以下のシミュレーションでは典型的な値として $0.8 < \beta < 1.2$ を用いる．

次に ③ の排重 (過程)，すなわち親兄弟と同じ名前を付けないというルールを以下のように実装した．以下に具体的な排重過程の例を示す．なお簡単のために $\beta = 1$ の場合を考えている．今，全ての祖先が表 2 の第 2 列のようなの名前分布であったとする．排重がない場合は，この第 2 列の名付け候補テーブルより，子の名は決定される．この場合，$\sum_j (s_j) = 50$ であるから，式 (2) の分母は 50 である．よって名 j が 1 になる確率は $s_1 = 8$ が分子となり，8/50 である．他の名 (j=2 10) になる確率も同様に計算できる．しかし排重過程がある場合には，第一子に関しては，親の名 (j=6 とする) が排重されるが，他の既存名については付けることが可能である．そのため第一子の名は，親の名のサイズを 0 にした名候補テーブル (表 2 の第 3 列) を用い，式 (2) に従って決まる．その結果，第一子の名が決まれば，さらにその名 (j=8 とする) のサイズを 0 にした名候補テーブル (表 2 の第 4 列) を用いて第二子の名を決定する．第三子以降に関しても同様である．

名 j	s_j	第一子	第二子
1	8	8	8
2	5	5	5
3	3	3	3
4	1	1	1
5	2	2	2
6	4	0	0
7	6	6	6
8	2	2	0
9	5	5	5
10	4	4	4
$\sum_j(s_j)$	50	46	44

表 2: 排重過程の説明. 名候補テーブル.

排重の影響を確かめるために，排重のない場合，親子・兄弟での排重がある場合のシミュレーションを行った (図 2). いずれの場合も，希少領域でべき則が見られた. また排重過程を加えることでべきの指数が小さくなり，ReaD で得られた指数 1.84 に近くなっていることがわかる.

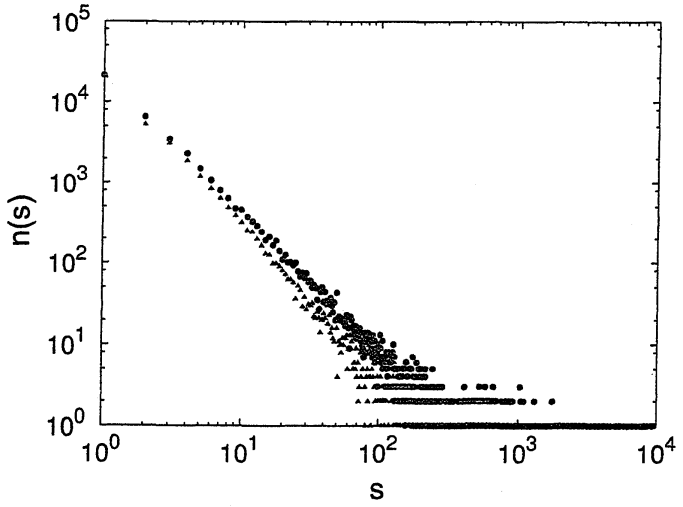

図 2: 赤丸は排重がある場合のサイズ頻度分布. べきの指数は 1.98. 青三角は排重がない場合. 指数は 1.81.

4 考察

実データとしてReaDからデータを取得し，姓・名のサイズ頻度分布を統計的に解析した．希少姓と稀少名のサイズと頻度がべき的な関係を持つことがわかった．またそれらのべきの指数の差は非常に小さい．姓と名ではその継承プロセスは異なるが，希少領域でのべき則の成立という同様の統計的特徴を示したことは興味深い．次に，Yule過程を取り入れた名付けモデルを立て数値計算を行った結果，希少名に関するべき則が得られた．また排重過程の有無により，べきの指数に変化が見られた．

姓・名がともに同様のべき則を示した理由としては，以下のような原因が考えられる．本研究のモデルはYule過程を取り入れた名付けに関するものであったが，Chida,Maseによる希少姓数の時間発展シミュレーションの結果では，希少姓は減少していくという．そうであれば，姓の分布がべきの指数を維持するとは考えにくい．そのため，姓と名が本研究で同様のべきの指数を示したのは，初期条件の影響と考えられる．つまり明治時代に日本国民の多くに姓が与えられた時に，Yule過程により選ばれた分布がべき的になっており，その影響が残っているために，同様のべきの指数を示したのではないか．したがって世代を経ればべきの値は変わるかもしれず，現在のべきは過渡状態なのかもしれない．

今後の課題としては，実データの時系列解析が挙げられる．ある年の姓・名の分布を初期条件として，モデルでシミュレーションを行い，その結果と1世代後(約25年後)もしくは1世代前の分布と比較を行うことで，モデルの妥当性や問題点を明らかできると考えられる．また，今回，姓と名を別々に扱ったが，本来姓と名はセットで個人を表しているので，姓・名セットのサイズ頻度分布も解析すべきである．そのためには，やはり大きな実データが必要である．

参考文献

[1] Y. Sato and H. Seno, 「姓の継承と絶滅の数理生態学」 京都大学学術出版会 (2003) .

[2] S. Miyazima, Y. Lee, T. Nagamine and H. Miyajima, "Power-law distribution of family names in Japanese societies", Physica A **278** (2000), 282–288.

[3] S. Miyazima, Y. Lee, T. Nagamine and H. Miyajima, "Family Name Distribution in Japanese Societies", J. of Phys. Soc. Jpn. **68** (1999), 3244–3247.

[4] S. Chida and S Mase, 「日本人の名字の統計解析」, 日本統計学会誌第**35**巻, 第1号 (2005) 55–70.

[5] ReaD研究開発支援総合ディレクトリ, http://read.jst.go.jp/.

[6] M.E.J. Newman, "Power laws, Pareto distributions and Zipf's law", Contemporary Physics **46** (2005), 323–351.

[7] R. Hayakawa, 「日本人の名前のサイズ頻度分布」大阪府立大学修士論文, (2012).