平成２３年度修士論文
日本人の名前のサイズ頻度分布

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
数理物理講座　非線形力学研究グループ
学籍番号 2100103042
早川良

2012 年 2 月

目 次

第 1 章　はじめに **3**

第 2 章　先行研究 **5**

2.1　統計量の定義 5

2.2　Miyazima らによる日本人の姓に関する実測結果 6

2.3　Baek らによる研究 10

第 3 章　実データの解析 **12**

3.1　ReaD 12

3.1.1　データの取得 12

3.1.2　姓に関する結果 14

3.1.3　名に関する結果 17

3.2　べきの指数の定義 21

3.3　累積頻度分布について 23

3.4　電話帳データ 24

3.4.1　データの取得 24

3.4.2　北海道のデータに関する解析結果 24

3.4.3　愛知県のデータに関する解析結果 26

3.5　希少姓 (名) 人口比率・希少姓 (名) 比率 30

第 4 章　モデル **31**

4.1　名付け過程のモデル化 31

4.2　世代と時間発展 33

4.3　結果 35

4.3.1　排重過程の影響 35

4.3.2　新名発生確率 α の影響 37

4.3.3　流行反映度 β の影響 38

4.3.4　α と β の影響 39

4.4　分布関数の時間発展 40

第 5 章　おわりに **42**

5.1　まとめ 42

5.2　今後の課題 42

謝辞 **43**

付録 ReaD のデータ取得方法 **47**

第1章　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、本論文では統一して “姓 (family name)” と呼ぶことにする。それに対し、名前を “名 (given name)” 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) も存在するので、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 らによるもの、Sato, Seno らによるものなどがある [1, 2, 3, 4]。Chida, Mase らは、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により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。外国の姓に関する研究としては、Baek らによるもの、Zanetti, Manrubia らによるもの、Scapori らによるものなどがある [5, 6, 7, 8]。その中で Baek らは、べき則を示す国とそうでない国があると主張している。

姓・名の分布は、時々刻々と変化している。その途中経過の一状態として、特徴的な分布が先行研究で見られたと考えられる。日本において分布に変化を与える要因は主に、誕生時の生成・婚姻 (離婚) 時の変更・死亡時の消滅の 3 つであると考えられる。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。

先行研究では、姓の分布に関するものがほとんどである。しかし、名についても多出名や希少名が存在することはまちがいない。我々は名の分布に着目し、その統計的な特徴に関する解析をした。データとしては、研究開発支援総合ディレクトリ

(ReaD)[9] と電話帳のデータを使用した。また、モデルを立てシミュレーションを行うことで、どういった名付けのメカニズムがべき則を成立させているのかを調べた。

本章の構成としては以下の通りである。第 2 章では、姓のサイズ頻度分布に関する先行研究を紹介する。第 3 章では、実データを用いて姓・名それぞれの統計分布を調査した結果を述べる。第 4 章では名付け過程のモデル化と数値計算の結果を紹介する。最後に、第 5 章でまとめと今後の課題について述べる。

第2章　先行研究

本章では先行研究を紹介する。まず 2.1 節で、統計量の定義を行い、2.2 節で日本人の姓に関する解析結果を紹介する。また、2.3 節では、幾つかの国の姓に関する解析結果を紹介する。

2.1　統計量の定義

本論文で扱う統計量を表 2.1 にまとめた。本節ではこれらについて解説していく。

統計量	記号	説明
総人口	S	データ内の人口
総姓数	N	データ内の姓の種類数
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
ランク	$r(s)$	サイズに対する順位

表 2.1: 本論文で扱う統計量。

総人口 S の人口集団に、異なる N 種類の姓があったとしよう。さらにその内訳は表 2.2 のように田中 30 人、佐藤 14 人、鈴木 14 人、伊藤 13 人…であったとする。そのとき、それぞれの姓のサイズ s は $s_{\text{田中}} = 30$, $s_{\text{佐藤}} = 14$ のようになる。頻度 $n(s)$ は、そのサイズを持つ姓の種類数であり、表 2.2 の場合は、サイズ 30 の姓は 1 つしかないので $n(30) = 1$, サイズ 14 の姓は 2 つあるので $n(14) = 2$ となる。頻度 $n(s)$ と総姓数 N、総人口 S の間には式 (2.1),(2.2) のような関係が成立する。

$$N = \sum_s n(s). \tag{2.1}$$

$$S = \sum_s sn(s). \tag{2.2}$$

統計量 / 姓	サイズ s(人数)	ランク r(s)	頻度 $n(s)$
田中	30	1	$n(30)=1$
佐藤	14	2	$n(14)=2$
鈴木	14	2	
伊藤	13	4	$n(13)=1$
⋮	⋮	⋮	⋮
希少姓	1	$N-n(1)+1$	$n(1)$
合計	S		N

表 2.2: 総姓数 N、人口 S の集団における統計量の例。

またランク $r(s)$ はサイズの関数であり、サイズの大きい姓から順に１，２，３…と与えられる。このとき、同じサイズの姓があればそれらには同じランクを与える。例では、サイズ 14 の姓が佐藤・鈴木の 2 つあるので、$r(14)=2$ となる。その次のサイズのものには、重複分を飛ばしたランクが与えられる。したがって、次にサイズの大きい姓は $s=13$ の伊藤であるが、重複分を考慮し $r(13)=4$ となる。よってサイズ s' の姓のランク $r(s')$ は式 (2.3) により決まる。

$$r(s')=\sum_{s'<s} n(s)+1 \tag{2.3}$$

2.2 Miyazima らによる日本人の姓に関する実測結果

Miyazima らは愛知県の電話帳データベースをもとに、5 つの地域の人口 S と総姓数 N を解析した。その結果が図 2.1 である。両対数プロットで、データが傾き 0.65 の直線上に乗っている。これより彼らは人口と総姓数の間にべき則

$$N \propto S^{\chi}, \quad \chi = 0.65 \pm 0.03, \tag{2.4}$$

が成立するとしている。

図 2.2(a) は姓のサイズと頻度の分布である。サイズを s/S^{α}, $\alpha = 0.37 \pm 0.03$ でスケーリングしたものが図 2.2(b) である。スケーリングすることで 5 つのデータが重なり、サイズの小さい領域でその傾きが一定になっていることがわかる。これより

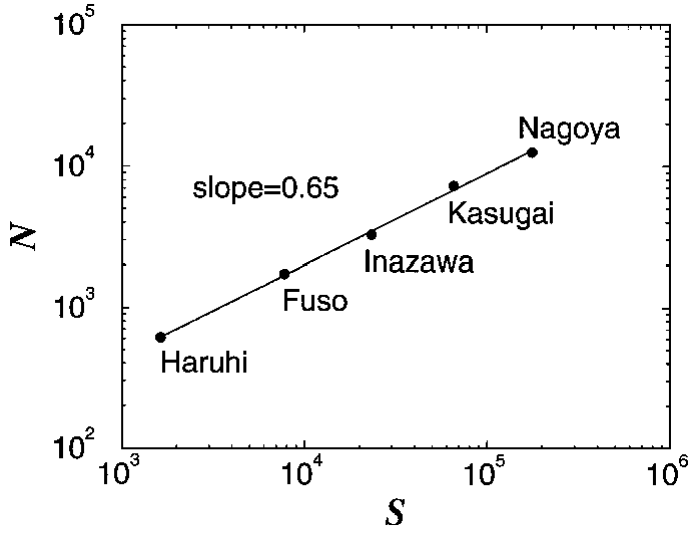

図 2.1: 5つの地域に対し、横軸に総人口 S, 縦軸に総姓数 N を両対数でプロットしたもの。傾きが 0.65 の直線に乗っている。(文献 [2] より転載)

彼らはべき則

$$n(s) \propto (s/S^{\alpha})^{-\gamma}, \tag{2.5}$$

を主張している。ただし、このとき $\gamma = 1.75 \pm 0.05$ である。

図 2.3(a) は姓のランクとサイズの分布である。これを $r/S^{\alpha'}$, $\alpha' = 0.5 \pm 0.05$ でスケーリングしたものが図 2.3(b) である。先ほどのサイズ頻度分布と同様に、こちらもスケーリングすることによって 5 つのデータが重なり、またクロスオーバー (折れ曲がり) が見られる。その前後で傾きが Φ_I から Φ_{II} と変化している。これより彼らは

$$s/S^{\alpha'} \propto \begin{cases} (r/S^{\alpha'})^{-\Phi_I}, \ r/r^* \ll 1 \\ (r/S^{\alpha'})^{-\Phi_{II}}, \ r/r^* \gg 1 \end{cases} \tag{2.6}$$

を主張している。ただしこのとき $\Phi_I = 0.67 \pm 0.03$, $\Phi_{II} = 1.33 \pm 0.03$ である。r^* とはクロスオーバーの起こるランク r の値で、$r^* \propto S^{\alpha'}$ という性質がある。また、クロスオーバーが現れる理由に関しては詳しく述べられていない。

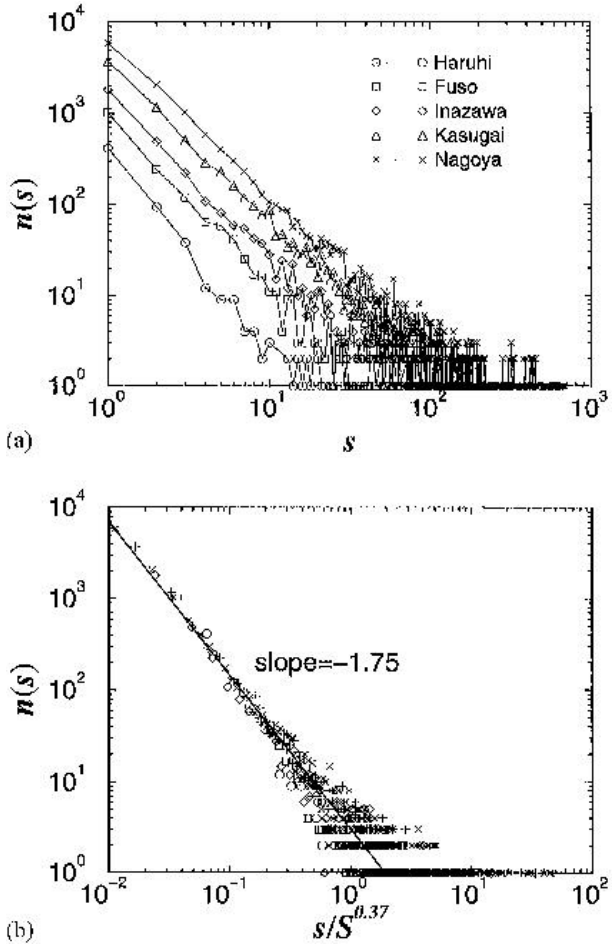

図 2.2: (a) 横軸にサイズ s, 縦軸に頻度 $n(s)$ を両対数でプロットしたもの。(b) 上のグラフを s/S^{α} でスケーリングしたもの。5 つのデータの希少領域が同一の傾き -1.75 になっている。(文献 [2] より転載)

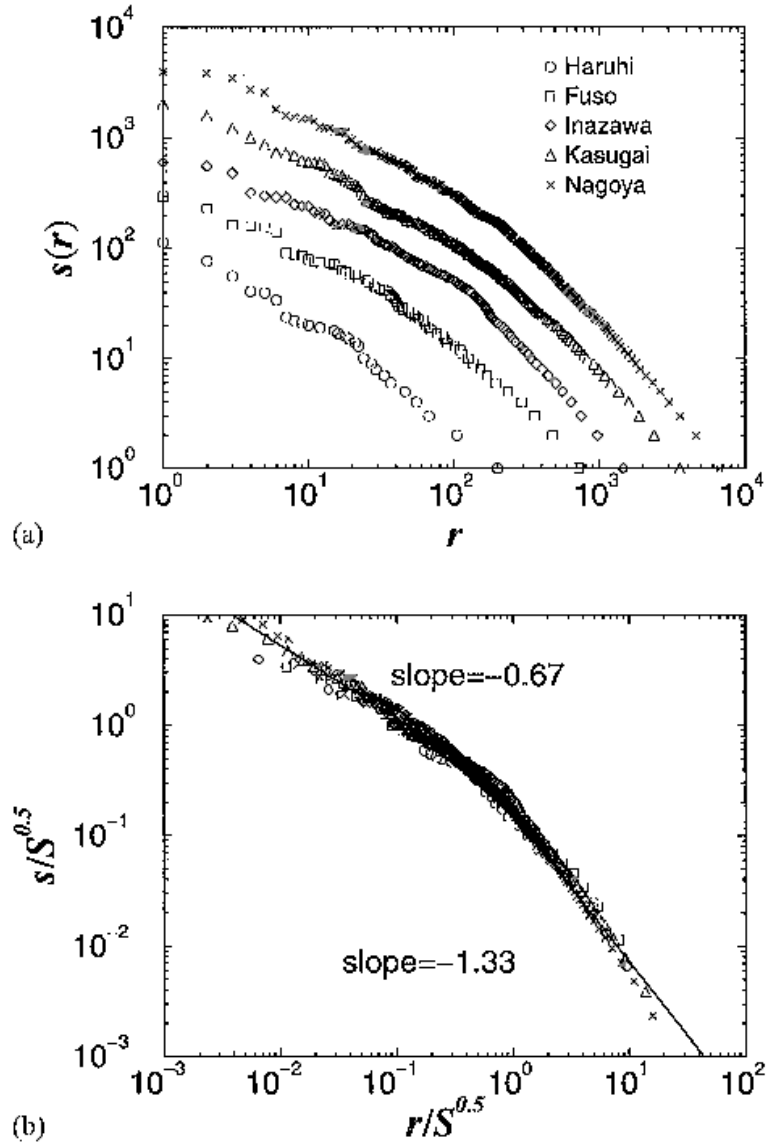

図 2.3: (a) 横軸にランク r, 縦軸にサイズ s を両対数でプロットしたもの。(b) 上のグラフを $r/S^{\alpha'}$ でスケーリングしたもの。ランクの大きな部分と小さな部分の間でクロスオーバーが見られる。その前後での傾きは -0.67 と -1.33 である。(文献 [2] より転載)

2.3 Baek らによる研究

Baek らの論文 [5] には、いくつかの国の姓のサイズ頻度分布に関する研究結果が紹介されており、それによると各国の希少姓のべきの指数は図 2.4 のとおりになっている。Baek らの主張によると、べき則を示す国は 2 つのグループに分けられ、べきの指数が 1.0 付近のものと、2.0 付近のものに分けられる。彼らは、新しい姓が出来るか否かによりそのグループ分けがされると主張している。$\gamma = 1.0$ 付近の国としては韓国があげられている。韓国は非常に姓の数が少なく、また新しい姓の誕生が制限されていた過去がある。その結果として、西暦 1500 年で 150 程度あった姓が西暦 2000 年でも 170 程度にしか増えていない (図 2.5)。それが原因で Korea のべきの指数は 1.0 となっていると主張している。

Region	γ	N_f
China [7]		$\ln N$
Korea [8]	1.0	$\ln N$
Argentina [9]		$N^{0.84}$
Austria [10]		$N^{0.83}$
Berlin [11]	2.0	
France [10]		$N^{0.90}$
Germany [10]		$N^{0.77}$
Isle of Man [12]	1.5	
Italy [10]		$N^{0.75}$
Japan [13]	1.75	$N^{0.65}$
Netherlands [10]		$N^{0.69}$
Norway [22]	2.16	
Sicily [14]	0.46–1.83	$N^{1.0}$
Spain [10]		$N^{0.81}$
Switzerland [10]		$N^{0.73}$
Taiwan [12]	1.9	
United States [11,15]	1.94	
Venezuela [16]		$N^{0.69}$
Vietnam [23]	1.43	$N^{0.27}$

図 2.4: いくつかの国の希少姓のべきの指数 γ。べきを示す国とそうでない国がある (文献 [5] より転載)。

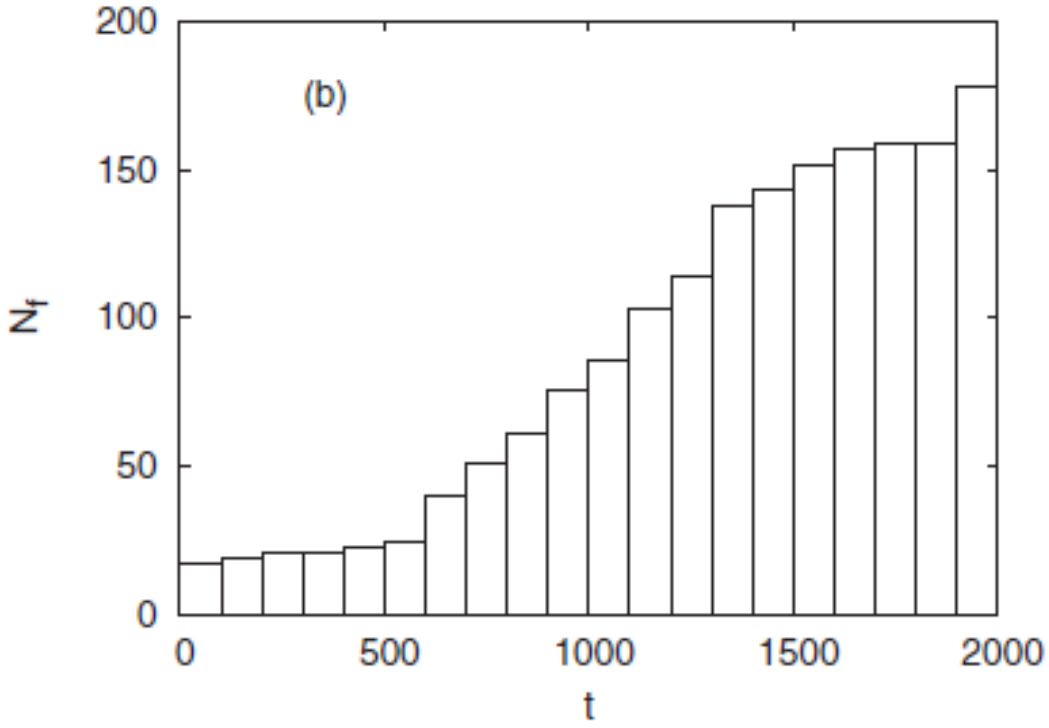

図 2.5: 西暦 t 年における韓国の総姓数 N_f の推移。1500 年以降に注目すると、500 年あまり間に姓がほとんど増えていないことが分かる (文献 [5] より転載)。

第3章　実データの解析

先行研究では解析の対象が姓であったが、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。姓と名を含む実データとして以下のものを用いた。

1. 研究開発支援総合ディレクトリ (ReaD)

2. 電話帳データ

1. は日本国内全体からのデータと考えられる。2. からは、全都道府県で最も登録件数の多い北海道と、先行研究 [2] との比較をするために愛知県のデータを選んだ。これらのデータの総人口 S、総姓数 $N^{姓}$ 及び総名数 $N^{名}$ の値を表 3.1 にまとめた。

統計量 / 地域	S	$N^{姓}$	$N^{名}$
ReaD	199,859	19,812	33,923
北海道	999,113	23,194	64,694
愛知県	955,412	24,069	74,781

表 3.1: 使用したデータの人口 S と総姓数 $N^{姓}$・総名数 $N^{名}$ の値。

本章では、これらの実データの解析結果を紹介する。

3.1　ReaD

3.1.1　データの取得

Miyazima らによる研究 [2] はデータが愛知県内のみと地域的に限定されており、全国の姓の分布をよく反映しているとは言い難い。すなわち、一つの地域として”Fuso”(愛知県丹羽郡扶桑町と思われる) を挙げており、そこでのランク 1 の姓が Senda(千田と思われる) でサイズが 296 ランク 2 の姓が Kondo(近藤と思われる) でサイズが 229 となっている (式 3.8)。

$$\begin{cases} s_{Senda} = 296, r(296) = 1 \\ s_{Kondo} = 229, r(229) = 2 \end{cases} \tag{3.1}$$

しかし、たとえば「日本の姓の全国データベース (以下、DB)[10]」を参照すると、千田は 476 位、近藤は 37 位であり、姓の上位ランクから外れている事がわかる (表 3.2)。

順位	ReaD(size)	DB	Fuso(size)
1	佐藤 (1893)	佐藤	Senda(296)
2	鈴木 (1699)	鈴木	Kondo(229)
3	田中 (1630)	高橋	
4	高橋 (1264)	田中	
5	伊藤 (1236)	渡辺	
6	中村 (1229)	伊藤	
7	山本 (1189)	山本	
8	小林 (1187)	中村	
9	加藤 (1044)	小林	
10	山田 (990)	加藤	

表 3.2: ReaD と DB[10] と Fuso の姓の上位ランクの比較

データベースの取得範囲を全国に広げれば、様々な地域特有の姓 (以下、地域姓と呼ぶ) が表れ、分布にも影響すると考えられる。そのため本研究では全国規模かつ豊富な姓・名のデータとして、研究開発支援総合ディレクトリ (ReaD) を利用した。得られたデータは 211,955 人分であった。なお取得方法の詳細は付録に記載した。解析の際、姓 (名) は漢字表記の差異で区別している。つまり、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、本研究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、次のような外国人姓と思われるものを除いた。

- 姓・名に全角半角のアルファベットを含むもの

- 姓・名に全角半角のカタカナを含むもの

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。

また、前述の DB と本研究で得た姓のランクの上位１０を比べると (表 3.2)、順位に差はあるが 9 つの姓が一致している。このことより、本研究のデータが全国的によく散らばっていることがわかる。

3.1.2 姓に関する結果

取得したデータについて姓のサイズ s、ランク $r(s)$、頻度 $n(s)$ を計算した。人口 S と総姓数 $N^{姓}$ を先行研究の図 2.1 に重ねたものが、図 3.1 である。ここで ReaD の点は先行研究の示す直線から少し上にずれているということに注目したい。つまり、ReaD の人口に対する総姓数が、Miyazima らによる解析に比べ多いということである。これに対して、以下のような要因が考えられる。本研究では、姓の区別を漢字表記で行ったが、Miyazima らがどのように区別したかは記載されていない。そのため、総姓数に差が見られた。また、ある特定の地域姓を多く含む先行研究のデータと比べ、ReaD のデータは様々な地域姓を含んでいるため、相対的に総姓数が多くなると考えられる。さらに、前節でも述べたが少数の外国人姓が残っているため、その影響で総姓数が増えているということも考えられる。

得られた姓のサイズ頻度分布を図 2.2(b) に重ねたものが図 3.2 である。先行研究のデータとうまく重なっており、およそ $s/S^{0.37} < 10^0$ の領域では、

$$n(s) \propto s^{-\gamma}, \tag{3.2}$$

の関係が成り立つことがわかる。サイズ s の小さい領域で線型近似をした結果、$\gamma = 1.83$ が得られた (べきの指数 γ の求め方については 3.2 参照)。

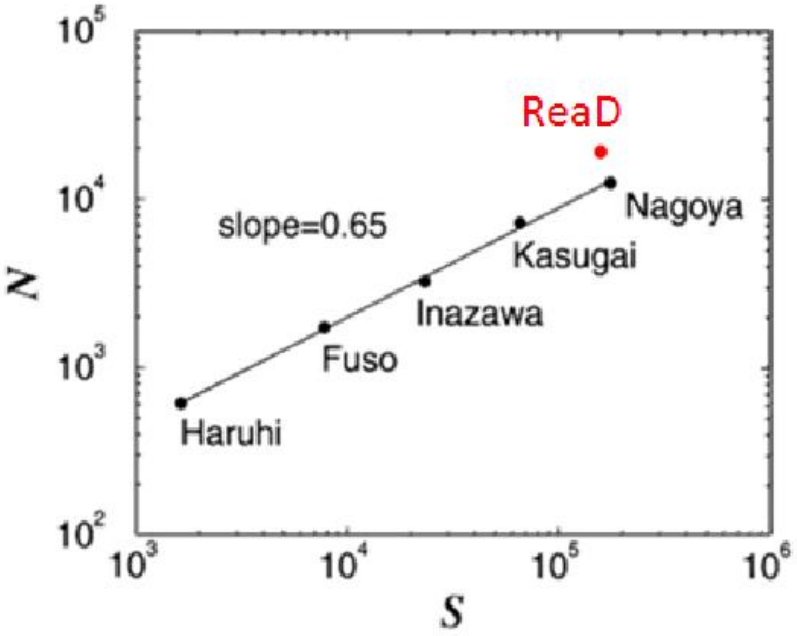

図 3.1: ReaD の人口 S と総姓数 $N^{姓}$ を図 2.1 に重ね合わせたもの。ReaD のデータは赤丸。

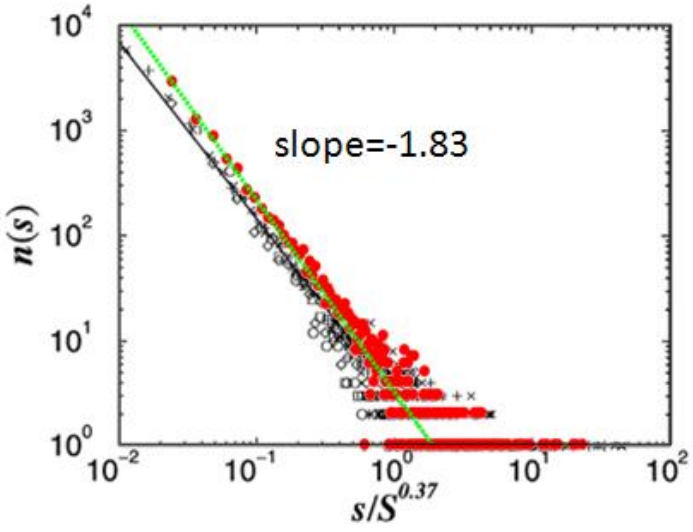

図 3.2: ReaD から得られた姓のサイズ頻度分布 (赤丸) を先行研究のグラフ (図 2.2(b)) に重ね合わせたもの。傾きが -1.83 の直線に乗っている。

得られた姓のランクサイズ分布を図 2.3(b) に重ね合わせたものが図 3.3 である。ReaD のデータは、ランクの上位 5 つくらいの点で先行研究の主張するべき則とずれていることが分かる。この原因の一つとしてもデータの地域性が考えられる。すなわち、先行研究の各地域でランク上位の地域姓の占める割合が高いからではないかと考えられる。

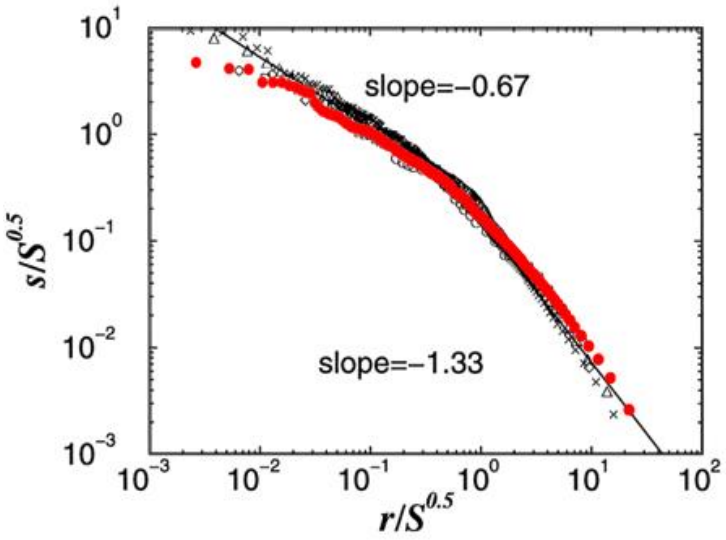

図 3.3: ReaD から得られた姓のランクサイズ分布 (赤丸) を先行研究の対応するグラフ (図 2.3(b)) に重ね合わせたもの。２つのべきと、それをつなぐクロスオーバーがみられる。

3.1.3　名に関する結果

姓と同様に、名に関しても解析した。得られた名の上位 10 位までを表 3.3 に示す。また、表 3.4 は明治安田生命が同社ホームページで公開している「生まれ年別名ベスト１０」[11] より生まれ年別名前ランキングを抜粋したものである。これらを比較すると一致する名が多数あることが分かる。ReaD に登録されているのは研究者であり、20 代後半〜60 才代の男性が多いと予想されるが、この予想と矛盾しない。

順位	名	サイズ
1	誠	733
2	徹	616
3	隆	588
4	健	510
5	淳	478
6	聡	475
7	博	470
8	修	462
9	浩	461
10	明	409
10	直樹	409

表 3.3: ReaD の名の順位とそのサイズ

姓と同様、名に関しても総名数 N_f、サイズ s、ランク $r(s)$、頻度 $n(s)$ を計算した。人口 S と総名数 N_f を図 3.1 に重ね合わせたものが図 3.4 である。総姓数と同様に先行研究の直線から大きく外れてはいないが、総姓数のおよそ 2 倍である。名は姓と違い任意に決められるという性質があるため、多様性があるのではないかと思われる。

順位＼生年 (西暦)	1960	1961	1962	1963	1964	1965
1	浩	浩	誠	誠	誠	誠
2	浩	浩	誠	誠	誠	誠
3	誠	浩一	豊	豊	修	修
4	浩二	徹	徹	修	隆	直樹
5	隆	剛	浩一	隆	達也	哲也
6	修	隆	修	浩一	豊	和彦
7	徹	和彦	和彦	和彦	和彦	豊
8	浩之	修	剛	哲也	直樹	剛
9	聡	浩二	隆	徹	浩一	学
10	博	聡	秀樹	直樹	勉	隆

順位＼生年 (西暦)	1966	1967	1968	1969	1970	1971
1	誠	誠	誠	誠	誠	誠
2	哲也	剛	大輔	大輔	大輔	大輔
3	剛	哲也	剛	学	直樹	健太郎
4	健一	直樹	健一	剛	剛	剛
5	学	健一	淳	大介	淳	大介
6	直樹	秀樹	哲也	直樹	大介	学
7	秀樹	学	直樹	健一	竜也	健一
8	徹	淳	学	淳	学	亮
9	英樹	英樹	聡	崇	健一	直樹
10	淳	大輔	大介	亮	亮	洋平

表 3.4: 生まれ年別名ランキング 1960〜1971 年。文献 [11] より転載。

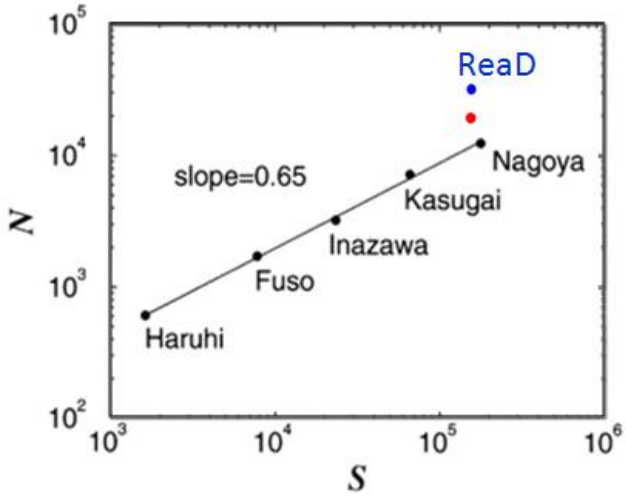

図 3.4: ReaD の人口 S と総名数を図 2.1 に重ね合わせたもの。ReaD の姓のデータは赤丸、名のデータを青丸で示している。

名のサイズ s と頻度 $n(s)$ のグラフを図 3.2 と重ねたものが図 3.5 である。$s/S^{0.37} < 10^0$ の領域で先行研究のデータとうまく重なっており、式 (3.2) の関係が成り立つことがわかる。また、フィッティングをした結果 $\gamma = 1.84$ が得られ、姓のサイズ頻度分布の $\gamma = 1.83$ と比べても傾きがほぼ一致することがわかる。ランク r とサイズ s のグラフを図 3.3 と重ねたものが図 3.6 である。傾きが姓のグラフとは異なり、はっきりした形でのクロスオーバーは見られない。以上より、サイズ頻度分布のべきの指数に関しては姓と名の相違は小さいが、ランクサイズ分布に関してはその傾きが大きく異なっている領域があることがわかる。

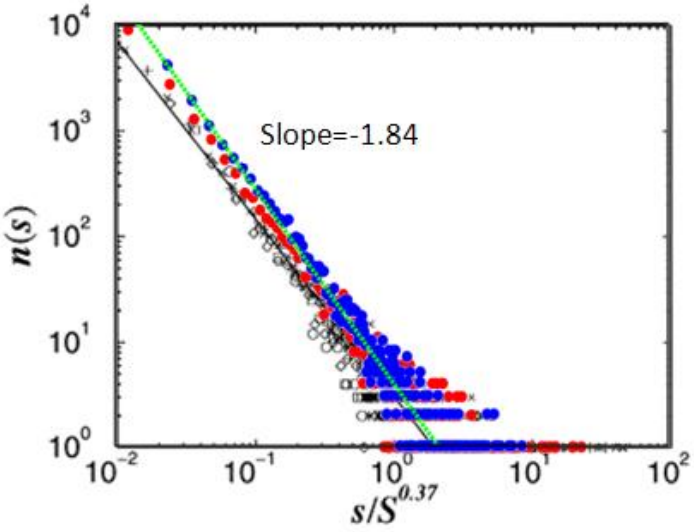

図 3.5: 名のサイズを横軸、頻度を縦軸に両対数でとったもの。赤丸が姓、青丸が名を示す。傾きが −1.84 の直線に乗っている。

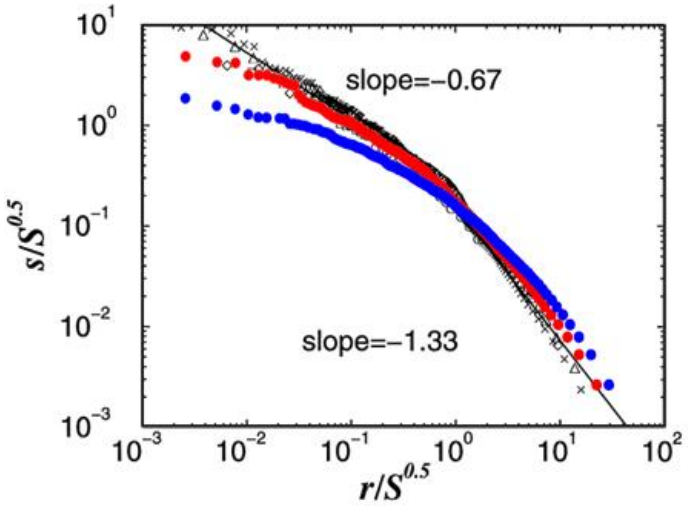

図 3.6: 名のランクを横軸、サイズを縦軸に両対数でとったもの。赤丸が姓、青丸が名を示す。はっきりした形でクロスオーバーが見られない。

3.2　べきの指数の定義

ある 2 変数 X, Y がべき則に従うことを

$$Y \propto X^{\gamma} \tag{3.3}$$

と表現する。サイズ頻度分布の典型的な例として、ReaD の姓のものを示す (図 3.7)。この図は両対数グラフであるが、サイズ s がおよそ 100 より小さい部分について、その分布が直線的になっていることがわかる。そのため、およそ $s < 100$ の領域で、s と $n(s)$ がべき則に従っていると言える。しかしどこまでがべき則に乗っているかを厳密に定義することができない。そのため、本研究では以下の通りにべきの指数を定義した。$n(s) = 1$ となる最小のサイズ s を s^* とし、両対数グラフにおいて、$s < s^*$ の領域について $y = ax + b$ で直線近似した。この際最小二乗法を用い、a の値をべきの指数としている。

$$n(s) \propto s^{\gamma}, \quad s < s^* \tag{3.4}$$

このときの γ をべきの指数と呼ぶ。ここで、べき則が成り立っている領域は $s < s^*$ であることに注意したい。この s^* を境に、$s < s^*$ の姓 (名) を希少姓 (名)、$s^* < s$ の姓 (名) を多出姓 (名) と呼ぶことにする。したがって、べき則は図 3.7 の希少姓 (名)

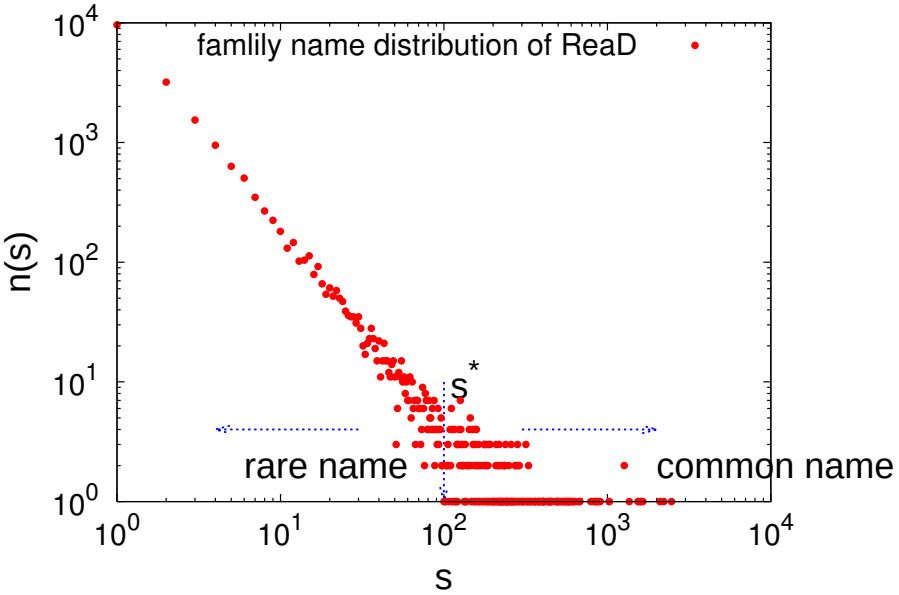

図 3.7: s^* および、希少姓 (rare name)・多出姓 (common name) の説明。

に関する統計的特徴だと言える。なお Miyazima らによる先行研究のべきの指数の測定方法については明記されていない。

次に、希少姓 (名) の割合を示す統計量を定義した。総人口に占める希少姓 (名) の人口比を r_S とし、

$$r_S = \frac{\sum_{s=1}^{s^*} sn(s)}{S} \tag{3.5}$$

で与えた。また総姓数 (総名数) に占める総希少姓 (名) 数の割合を r_N とし、

$$r_N = \frac{\sum_{s=1}^{s^*} n(s)}{N} \tag{3.6}$$

で与えた。

3.3 累積頻度分布について

希少領域に対する特徴を見るときには累積頻度分布を見るという方法もある。累積頻度 $c(s')$ とはサイズ s' 以上の姓の数が、総姓数 N に占める割合であるから、

$$c(s') = \frac{\sum_{s' \leq s} n(s)}{N} \tag{3.7}$$

で与えられる。例として、ReaD の姓と名の累積頻度分布を計算した (図 3.8)。姓に関してはおよそ $s < 200$、名に関してはおよそ $s < 60$ の領域で直線上に乗っている。サイズ頻度分布のべきの指数が γ であるとき、累積頻度分布のべきの指数は $\gamma + 1$ になることが知られている。サイズ頻度分布における傾き γ は姓が -1.83、名が -1.84 であったので、それらの値に 1 を加えた傾きの直線を図に載せている。図を見ると、直線と分布が一致していることがわかる。

サイズ頻度分布ではべき則の境界として特徴的なサイズ s^* が見られるのに対し、累積頻度分布ではそのような特徴的なサイズがわかりにくく、どこまでがべき則に乗っているのかが不明瞭である。したがって本論文ではサイズ頻度分布に注目し、姓・名の統計的特徴を見ている。

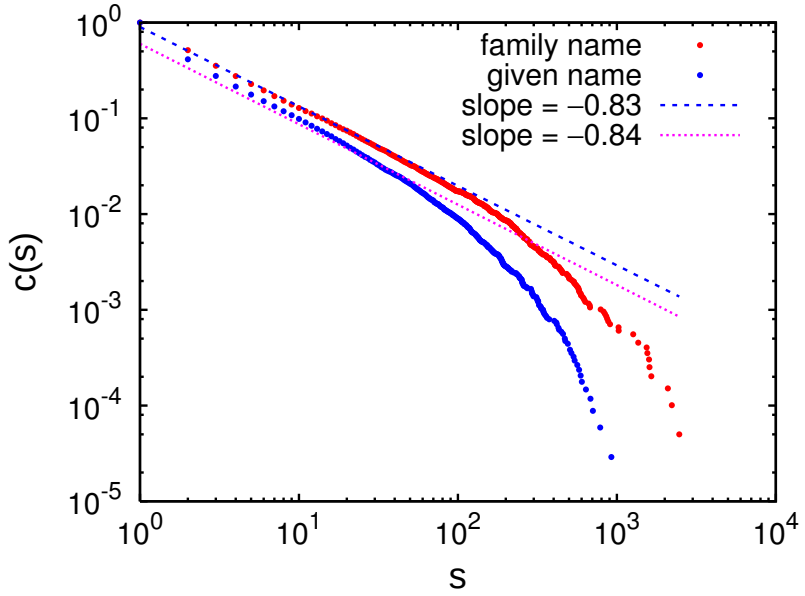

図 3.8: ReaD の姓・名のサイズと累積頻度分布。

3.4 電話帳データ

3.4.1 データの取得

ReaD とは別に電話帳 CSV 販売.jp のデータに対する解析も行った。電話帳からは、各都道府県のから 2 種を選択した。1 つ目は、全ての都道府県の中で登録人数が最も多い北海道を選択した。2 つ目は、登録人数も多く、Miyazima らの研究と比較が可能なことから、愛知県を選択した。データ形式は、

電話番号・姓　名・郵便番号・都道府県・市区町村・番地

となっており、2 番目の姓と名の部分を取り出して利用した。なお元データと思われる電話帳（タウンページ）には、このように姓と名の間に全角スペースはなく、姓と名の境目が特定できない (例えば 大阪　太郎 と 大　阪太郎)。しかし本データでは何らかの特定方法により姓と名の間には全角スペースが入っていたので、それを信頼し解析に利用した。

3.4.2 北海道のデータに関する解析結果

北海道の電話帳データは登録件数 999, 113 件である。サイズ頻度分布を図 3.9 に、ランクサイズ分布を図 3.10 に示す。図 3.9 を見ると、希少名については直線に乗っていることがわかる。しかし希少姓については $10 < s < s^*$ の領域では直線上に乗っているが、$s < 10$ の領域で直線から下に外れている部分がある。すなわち北海道には、サイズの極めて小さい (珍しい) 姓が比較的少ないということである。このような傾向は Chida, Mase らによる研究でも見られる。

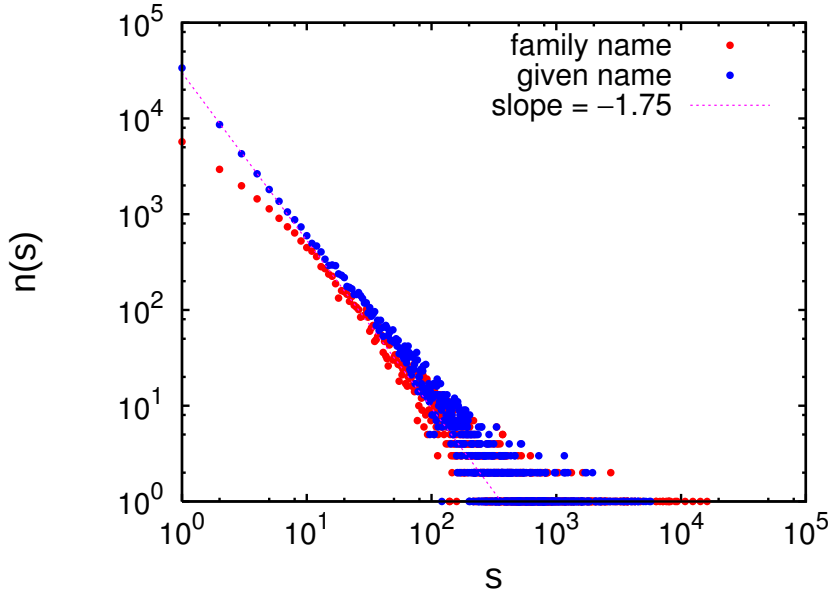

図 3.9: 北海道の姓と名のサイズ頻度分布。サイズが 1〜10 程度の姓について、べき則から下に外れているが、名についてはそのようなの傾向は見られない。

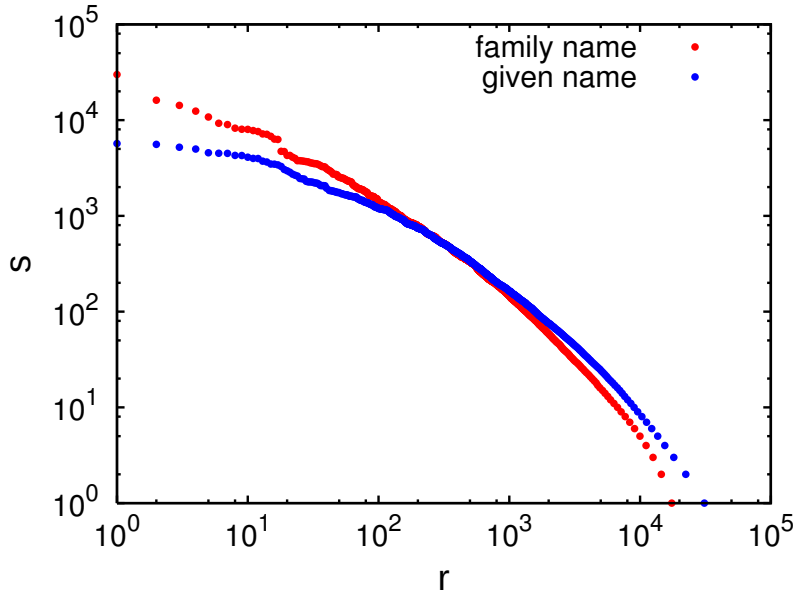

図 3.10: 北海道の姓と名のランクサイズ分布。明確なべき則やクロスオーバーは見られない。

3.4.3 愛知県のデータに関する解析結果

愛知県の電話帳データは登録件数 995, 412 件である。サイズ頻度分布を図 3.11 に示す。姓に関しては、北海道の姓 (図 3.9) の $s < 10$ の領域で見られたような、べき則からの落ち込みは見られない。また名に関して、よくべき則に乗っている事がわかる。以上のデータから、希少名のサイズ頻度のべき依存性について、地域的な差異は認めにくく普遍的な性質と期待される。図 3.12 に示したランクサイズ分布から、姓に関してのクロスオーバーが見られた。

先行研究における Fuso は丹羽郡の扶桑町だと思われるが、その根拠を以下に述べる。Miyazima らの論文 [2] には、ランク $1, 2$ の姓とそのサイズが例示されており $s_{Senda} = 296$、$s_{Kondo} = 229$ である (式 3.8)。本解析でも姓のランク $1, 2$ の姓は同一で、

$$\begin{cases} s_{千田} = 212, r(212) = 1 \\ s_{近藤} = 155, r(155) = 2 \end{cases} \tag{3.8}$$

であった。これより人口に占める千田と近藤の割合を計算すると、千田は 3.81%から 4.09%，近藤は 2.95%から 2.99% と、あまり変化していないことが分かる。

また、先行研究は 2000 年に発表されたものであるから、その解析データはそれ以前のものである。今回使用したデータは 2011 年のものなので、いくつかの統計量の時間変化が見られる。先行研究での調査対象であった愛知県内の 5 つの地域の人口、及び本解析で得られた人口、姓・名の γ を表 3.5 にまとめた。人口に大きな変化が見られるが、これは市区町村の合併や分割や、電話帳に非掲載の世帯が増えているためと考えられる。

図 3.13 では、5 つの地域のサイズ頻度分布 (2011) である。対応する図としては、図 2.2(a) である。春日や扶桑などの分布は、直線上に乗っているとは言いがたい。特に春日は $s^* = 9$ であり、べきに則に乗っている部分が 1 桁足らずである。これを先行研究と同様にスケーリングして得られたものが図 3.14 である。対応する図としては、図 2.2(b) である。スケーリングすることで、５つのデータが重なり、分布が同一直線上に乗っているように見える。さらに、これらの 5 つの地域をひとまとめにしたデータのサイズ頻度分布を見てみた (図 3.15)。これにより $\gamma = 1.65$ が得られた。

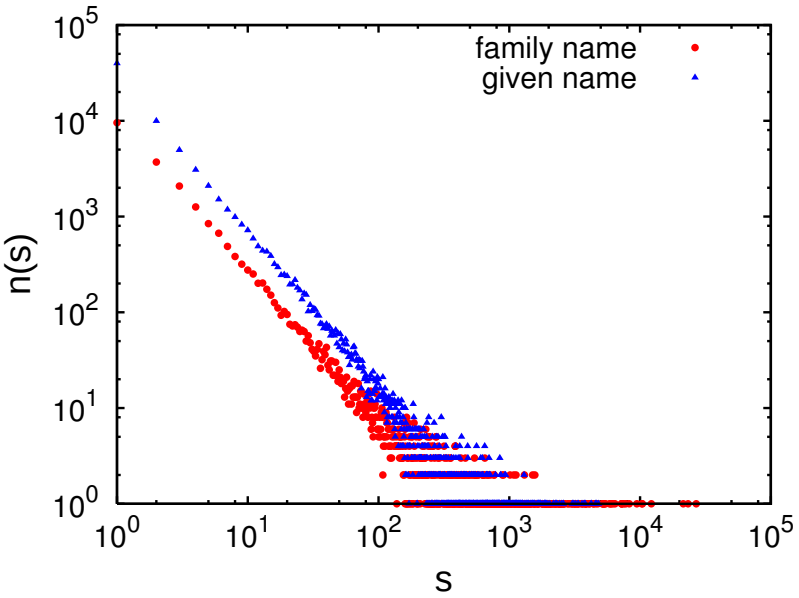

図 3.11: 愛知の姓と名のサイズ頻度分布。姓・名ともによくべき則に乗っている。

	Miyazima(2000)	本研究 (2011)		
地域	S		$\gamma^{姓}$	$\gamma^{名}$
Haruhi	1,634	988	2.23	3.24
Fuso	7,775	5,178	2.16	2.39
Inazawa	23,365	21,252	1.83	2.16
Kasugai	65,988	39,649	1.79	2.02
Nagoya	177,267	238,673	1.68	1.74
5region	276,029	305,750	1.65	1.85
Aichi		995,412	1.60	1.74

表 3.5: Miyazima らの研究 [2] との比較。Haruhi は清須市の春日、Fuso は丹羽郡扶桑町、Inazawa は稲沢市、Kasugai は春日井市、Nagoya は名古屋市全域のものと対応付けた。市町村合併に伴うと思われる人口サイズの変化が見られる。

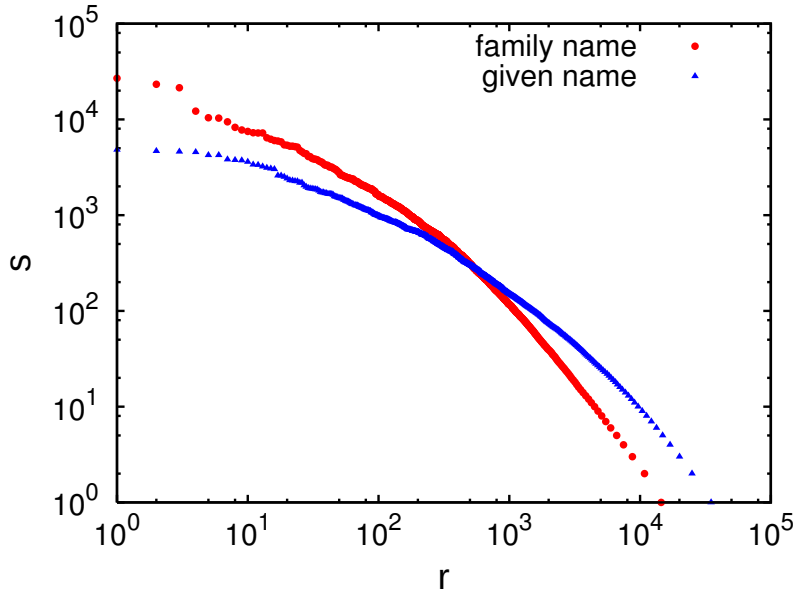

図 3.12: 愛知の姓と名のランクサイズ分布。明確なべき則やクロスオーバーは見られない。

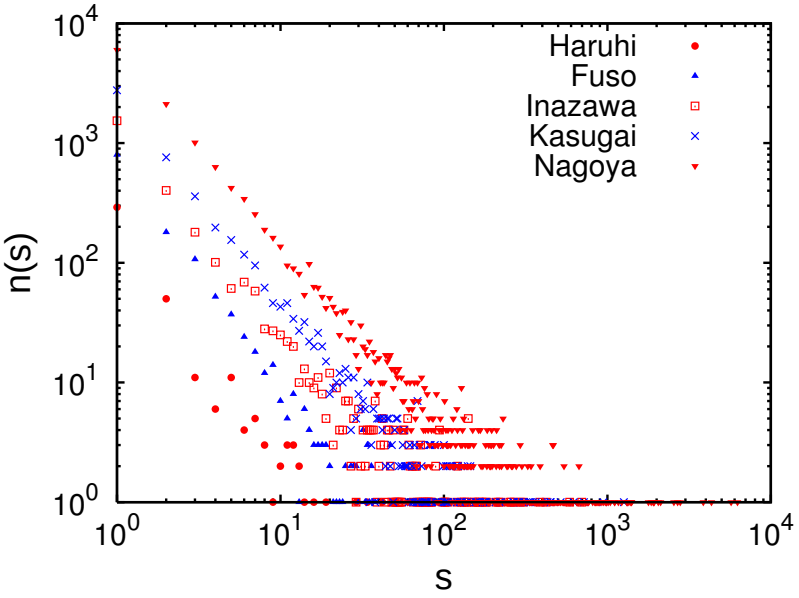

図 3.13: 清須市春日、丹羽郡扶桑町、稲沢市、春日井市、名古屋市全域のサイズ頻度分布。

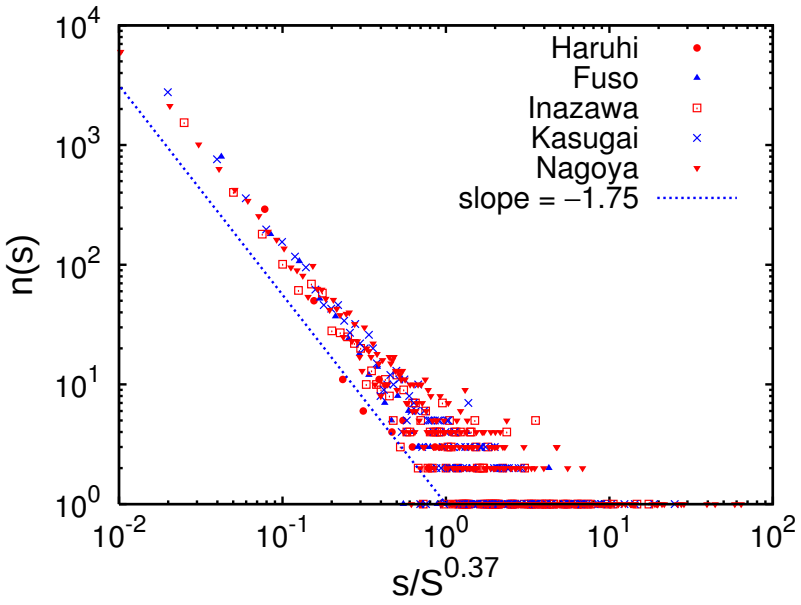

図 3.14: 清須市春日、丹羽郡扶桑町、稲沢市、春日井市、名古屋市全域のサイズ頻度分布をスケーリングしたもの。

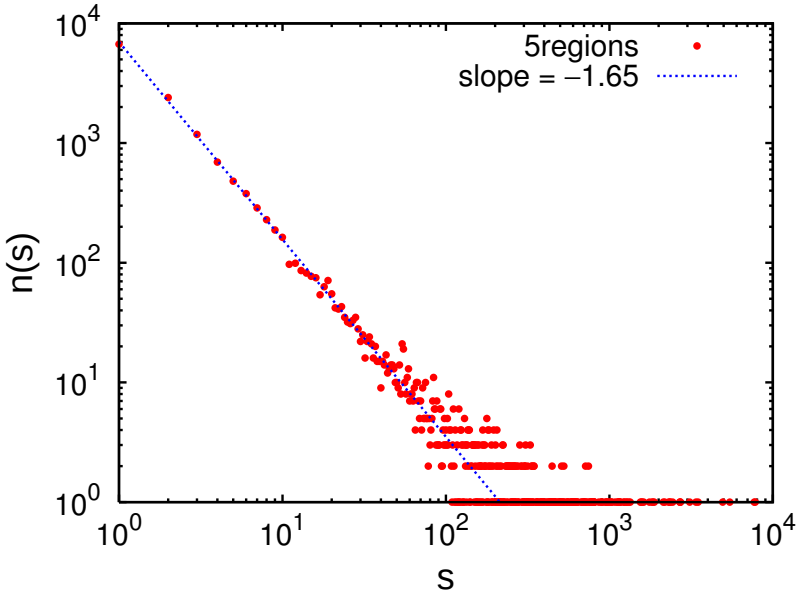

図 3.15: 清須市春日、丹羽郡扶桑町、稲沢市、春日井市、名古屋市の５つのデータをひとまとめにしたもののサイズ頻度分布。$\gamma = 1.65$ が得られた。

3.5　希少姓 (名) 人口比率・希少姓 (名) 比率

本節で解析した各データの希少姓比率 $r_N^{姓}$ と希少姓人口比率 $r_S^{姓}$、および希少名比率 $r_N^{名}$ と希少名人口比率 $r_S^{名}$ の値を表 3.6 と図 3.16 にまとめた。人口サイズ S によらず、r_N は 1 に近い値をとっており、総姓数 (総名数) に占める希少姓 (名) の割合がかなり大きいことが分かる。一方、S が大きくなるに従って r_S は小さくなっていることが見て取れる。

	$r_N^{姓}$	$r_S^{姓}$	$r_N^{名}$	$r_S^{名}$
ReaD	0.983	0.490	0.987	0.637
北海道	0.956	0.255	0.979	0.378
愛知県	0.963	0.189	0.988	0.489

表 3.6: ReaD・愛知県・北海道のデータにおける r_N, r_S の値。

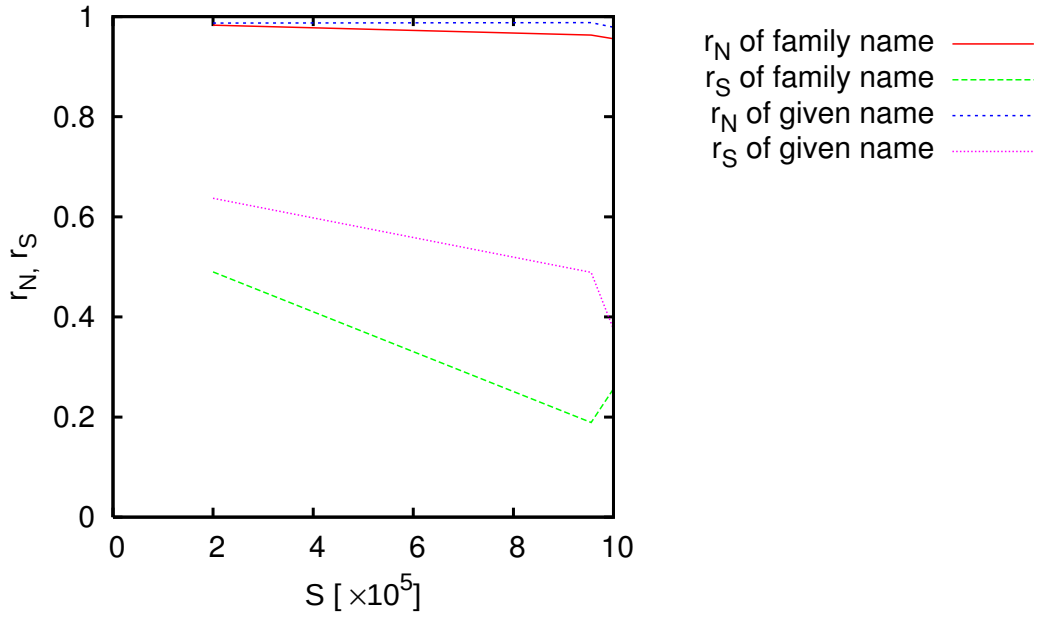

図 3.16: ReaD・愛知県・北海道のデータに関する人口 S と r_N, r_S の比較。

第4章　モデル

第3章での実データ解析により、希少姓と希少名それぞれのサイズと頻度の間にべき則があるとわかった。姓・名の分布は、時々刻々と変化しているが、その途中経過もしくは定常状態として、希少領域でべき則が観測されたと思われる。日本において分布に変化を与える要因は主に、誕生時の生成、婚姻 (離婚) 時の変更、死亡時の消滅の3つであると考えられる。姓に関しては基本的に父親の物を受け継ぐというメカニズムがあるが、名に関しては、幾つかの制約はあるが基本的に任意に決める事が可能である。そのため姓はその数が増えることは稀で、減少傾向にあると考えられる。しかし名についてはよくわかっていない。このように性質の違いのある姓と名が、その希少領域でべき則を示し、またその指数が似通っていることは非常に興味深い。

姓の継承に対しては、Sato, Seno らが Galton-Watson モデルを用いた数値計算を行なっている [1]。そこで本研究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考えていく。それらのパラメータを設定することで、どのようなメカニズムが希少名のサイズ頻度分布に影響を与えるのかを調べていく。

4.1　名付け過程のモデル化

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 [12] 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。模式的に書くと図 4.1 のようになる。

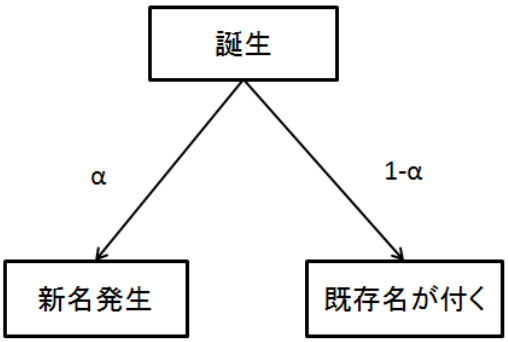

図 4.1: 個体が誕生したときに新名が付くか、既存名が付くかの模式図

この時にどの既存名が付くかは、以下のように決まる。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 $P(j)$ は、

$$P(j) = \frac{s_j}{\sum_j s_j} \tag{4.1}$$

で与えられる。必ずいずれかの既存名が付くことから

$$\sum_j P(j) = 1 \tag{4.2}$$

が成立する。これは Yule 過程に他ならない。今回、さらにこの過程を一般化し、確率 $P(j)$ を

$$P(j) = \frac{(s_j)^\beta}{\sum_j (s_j)^\beta} \tag{4.3}$$

と与える。この β は実数パラメータである。例えば $\beta = 0$ では、すべての名が同一の確率 $\frac{1}{N}$ で選ばれ、$0 < \beta < 1$ では、$P(j)$ のサイズ依存度が小さくなる。$\beta = 1$ では式 4.1 に示した単純な Yule 過程となり、$1 < \beta$ では、$P(j)$ のサイズ依存度が大きくなる。以下では、典型的な値として［0.8,1.2］の β を用いる。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るから、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 β を流行反映度と呼ぶことにする。

次に ③ の排重 (過程)、すなわち親兄弟と同じ名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 4.1 の第 2 列のようなの名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 (4.3) 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、$\frac{8}{50} = \frac{4}{25}$ 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 4.1 の第 3 列) を用い、式 (4.3) 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 4.1 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 4.1: 排重過程の説明。名候補テーブル。

4.2 世代と時間発展

本モデルでは、名付けプロセスに排重過程を取り入れるため、個体の親子関係を明確にする必要がある。そこで “世代” を定義した。一般に各個体が子供を産むまでの年数は異なるため、世代の重複が起こりうる。たとえば図 4.2 のような家系図の場合で、個体 α の高祖父 ① と曽祖母 ② が同じ時代に生まれており、世代の重複が起きている。しかし今回はモデルの単純化のために、ある期間が経つとすべての個体は一斉に次の世代を作るものとする。このようにすることで世代の重複をなくしてい

る (図 4.3)。次に、次世代の作り方を説明する。ある世代の個体全てに子を作るが、子の数は平均が λ のポアソン分布に従うものとした。このとき子は先ほどの過程で名付けられ、その個体群が親となりまた子を作る。このように世代交代を繰り返すことで時間発展させる。ある世代 G の人口を $S(G)$ とし、一定値を越えるまで世代交代を繰り返す。

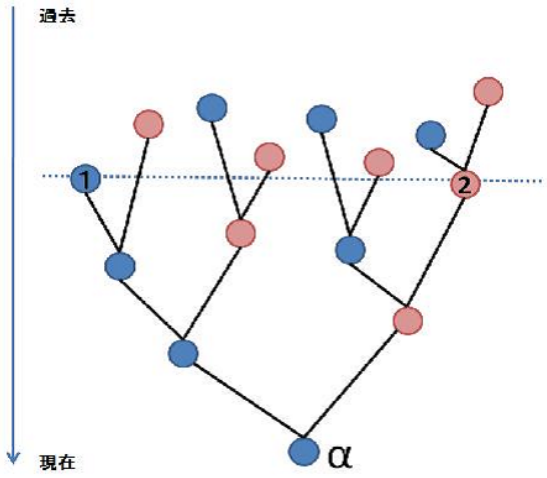

図 4.2: 世代の重複がある。個体 ① と個体 ② は同じ時代に生まれているが、個体 α にとっては高祖父と曽祖母であり、世代が重複している。

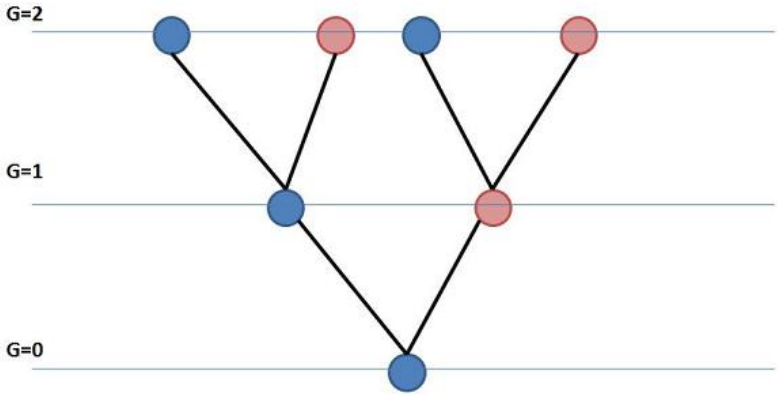

図 4.3: 一定の期間が経つと一斉に次世代が生まれる。そのため世代の重複がない。現実の世界とは異なる。

以降紹介するシミュレーションでは、特に指定しない限りデフォルトの初期条件・パラメータとして以下の表 4.2 の値をとるものとする。この初期条件分布 $n(3) = 3$ とは、$\{s_1, s_2, s_3\}$=$\{3,3,3\}$ と同義であり、総名数は $N = 3$ で総人口は $S = 9$ である。

パラメータ	値
α	0.001
β	1.0
初期条件分布	$n(3) = 3$
λ	1.3
排重過程	親兄弟
終了条件	$N(G) \geq 8 \times 10^6$

表 4.2: パラメータのデフォルト値。

4.3 結果

4.3.1 排重過程の影響

排重の影響を確かめるために、[排重のない場合] [親子の排重のみがある場合] [兄弟の排重のみがある場合] [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。いずれの場合も、希少領域でべき則が見られた。典型的な例として、排重のない場合 (non-exclusive) と親子・兄弟の排重のある場合 (exclusive) のサイズ頻度分布を図 4.4 に示した。排重の無い場合に比べ、排重のある場合の方が、希少領域の傾きが小さくなっていることが見て取れる。表 4.3 に、4 つのケースでのべきの指数を示す。排重過程を加えることでべきの指数が小さくなっており、その数値は ReaD や北海道の電話帳データで見られた指数に近くなっていることがわかる。以上のことから、排重はべきの指数を小さくする影響を及ぼすということがわかった。以後、排重の有無についての記述があるが、排重あり (exclusive) とは親・兄弟の排重がある場合のことを示す。

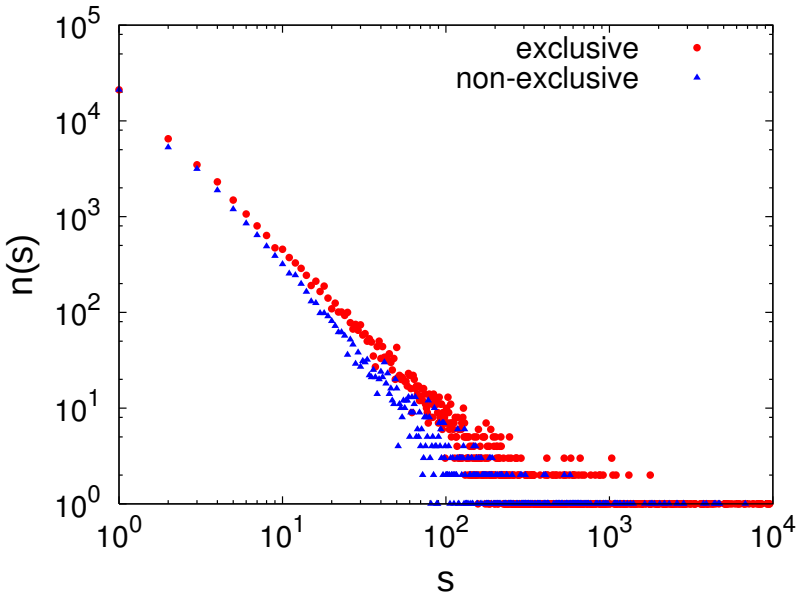

図 4.4: 排重過程の有無によるサイズ頻度分布の比較。赤丸は排重あり、青三角は排重なしを示す。排重がない場合に比べ、排重がある場合のほうが傾きが緩やかになっていることがわかる。

排重	γ
なし	1.98
親子のみ	1.82
兄弟のみ	1.86
親・兄弟	1.81

表 4.3: 各排重過程の γ への影響。

4.3.2 新名発生確率 α の影響

新名発生確率 α の値を 0.00001〜0.1 まで変化させて数値計算した。その結果、サイズ頻度分布の希少名領域にべき則が見られた。この時の α とべきの指数 γ の関係を図 4.5 に示した。なお、各 α の値に対し 10 回のシミュレーションを行い、べきの指数の平均と標準偏差を表示している。これを見ると、α の増加に伴い、べきの指数 γ が増加している事がわかる。また、排重の有無で γ の α 依存性が変化している事がわかる。

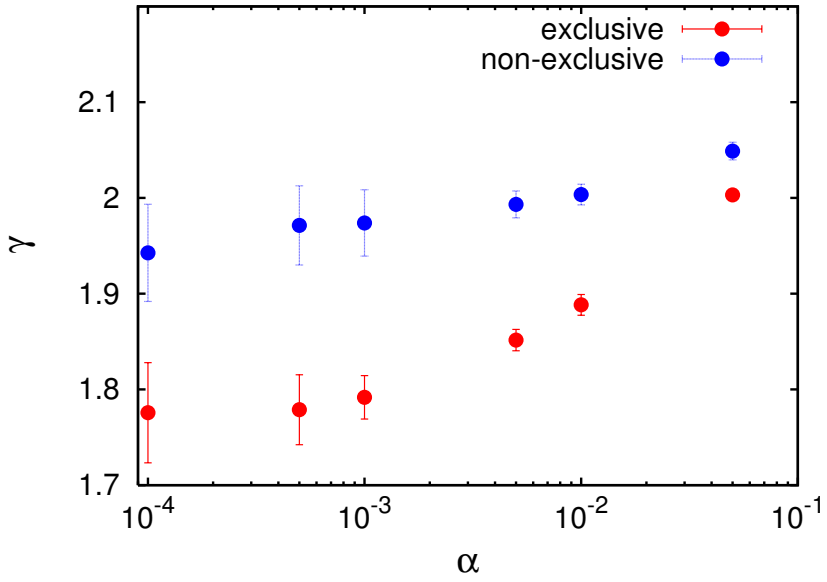

図 4.5: α を変化させた時 γ の値の変化。赤丸 (青丸) は排重のある (ない) 場合の 10 回のシミュレーションの γ の平均値、エラーバーはその標準偏差を示す。

4.3.3 流行反映度βの影響

次に、流行反映度 β の値を 0〜1.3 まで 0.05 刻みで変化させて数値計算した。$0.8 \leq \beta$ では、サイズ頻度分布の希少領域についてべき則が成り立った。例として、$\beta = 0.8, 1.15$ の排重のある場合のサイズ頻度分布を図 4.6 に示す。べき則が成り立つ場合 $(0.8 \leq \beta)$ の β と γ の関係を図 4.7 に示す。β の増加に伴い、γ も増加していることがわかる。

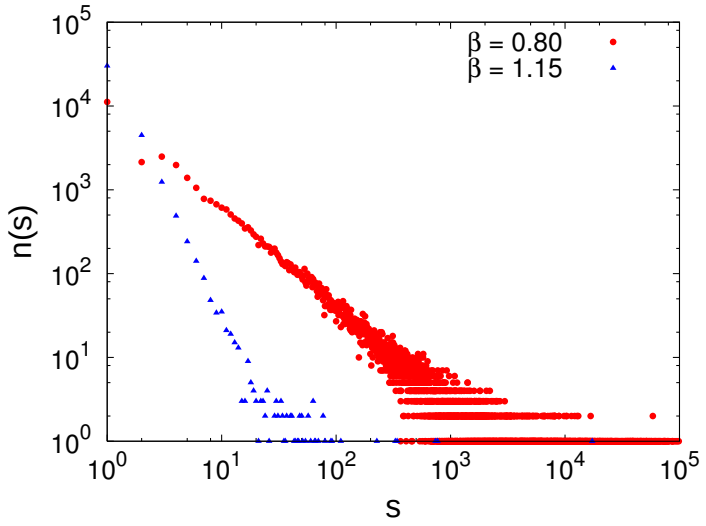

図 4.6: $\beta = 0.80$(赤丸)、$\beta = 1.15$(青三角) のときのサイズ頻度分布。

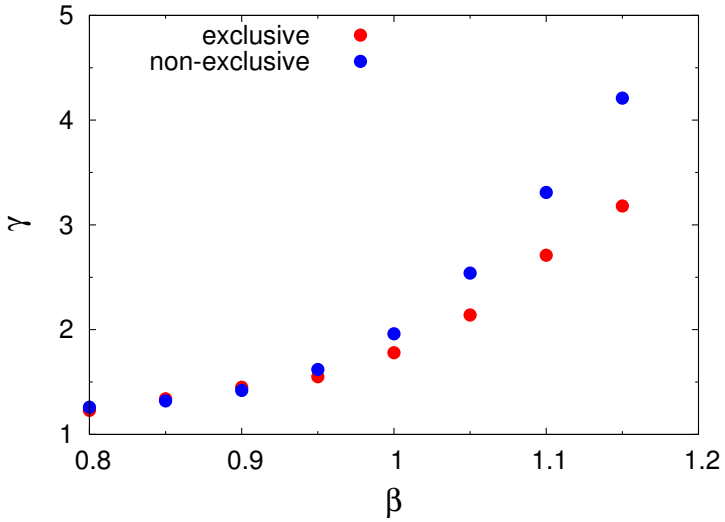

図 4.7: β を変化させた時の指数 γ の変化。赤丸は排重あり、青丸は排重なし。

$\beta < 0.8$ では、サイズ頻度分布がべき的にならず、波が幾重にも重なったような構造を示した。例として $\beta = 0.70$ のときのサイズ頻度分布を図 4.8 に載せる。

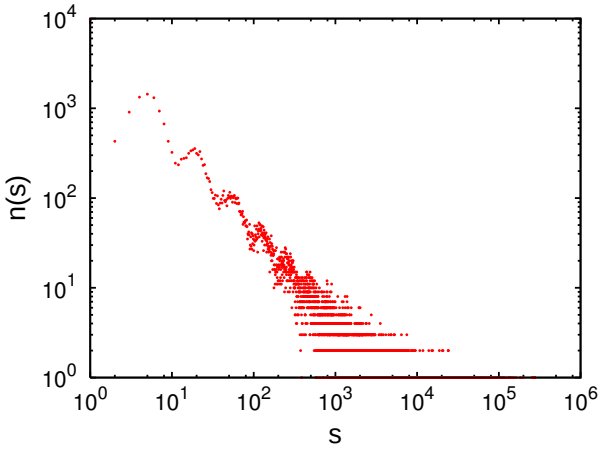

図 4.8: $\beta = 0.70$ のサイズ頻度分布。直線的でなく波状になっている。

4.3.4 αとβの影響

以上の結果より、新名発生確率 α や流行反映度 β、排重過程などにより希少領域のべきの指数が変化することがわかった。しかし実際にはそれらの影響の複合によって、べきの指数は決定されていると思われる。そこで、α と β をともに変化させ、γ の値を計測した (図 4.9)。α と β により、べきの指数 γ が変化しており、$(0.0001 < \alpha < 0.05, \beta = 1.0)$ 付近で、実データから得られる $\gamma = 1.8$ に近い値がみられている。

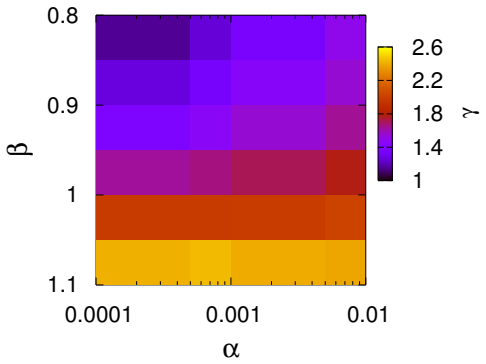

図 4.9: γ の α, β 依存度

4.4 分布関数の時間発展

本節ではシミュレーションで得られた分布の時間発展について述べる。実データ解析で、姓・名はともに、そのサイズ頻度分布がべき則を示すことがわかった。しかし、今後時間発展することで、そのべきは維持されるのだろうか。またべきが維持される場合、べきの指数がどのように変化していくのだろうか。これらの問題に対して我々はモデルによって得られた分布の時間発展を調べた。

図 4.10 に世代ごとのサイズ頻度分布を示した。$G = 40$ 以降で、希少領域にべき則が見られる。また世代が大きくなるにつれ s^* の値は大きくなり、べき則に乗っている桁数が増えていることがわかる。

図 4.11 には、べき則が見られ始めた G=40 以降の γ の値の変化を載せている。$G = 50$ 以降で γ の振れ幅がかなり小さくなっており、概ね $\gamma = 1.8 \pm 0.02$ に収まっているといえる。このことから、この例では $G = 50$ で $\gamma = 1.8$ に収束していると見なす。

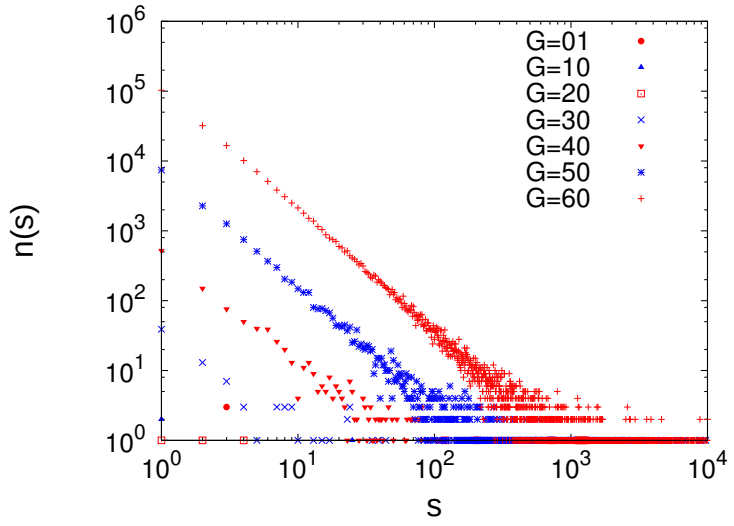

図 4.10: G=1, 10, 20, 30, 40, 50, 60 におけるサイズ頻度分布。

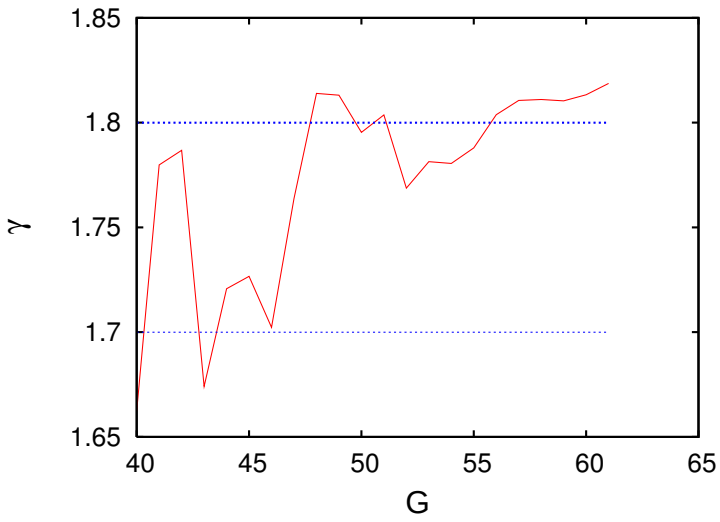

図 4.11: 世代発展に伴うべきの指数の変化。

第5章　おわりに

5.1　まとめ

実データとしてReaD、電話帳 (北海道・愛知県) からデータを取得し、姓・名のサイズ頻度分布とランクサイズ分布を統計的に解析した。希少名に関してはすべてのデータでべき則が見られた。このことから名の分布の地域的普遍性が示唆される。希少姓に関しても、概ねべき則が成り立っていた。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。

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、新名発生確率α、流行反映度βにより、そのべきの指数に変化が見られた。またα, βの組み合わせで、実データで観測されたべきの指数に近いものを得ることに成功した。

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ら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。よって世代を経ればべきの値は変わるかもしれず、現在のべきはその途中経過かもしれない。

5.2　今後の課題

今後の課題としては、時系列を伴った実データを得たい。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にしたい。

今回、姓と名を別々に扱ったが、本来は姓・名の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見てみたい。そのためには、やはり大きな実データが必要である。

本研究ではサイズ頻度分布の、特に希少名のべきの指数に注目して、実データとモデル結果の比較を行ったが、r_S, r_N の値は大きく異なっている。そのため、多出名の統計的特徴をうまく捉えられていないモデルだと考えられる。より現実世界に即したモデルを構築することが課題として残っている。

謝辞

本研究を行うにあたり、様々な方にお世話になりました。細部に渡り御指導いただきました水口毅講師には、深く感謝致します。大同寛明教授、堀田武彦准教授、福田浩昭助教には様々なアドバイスをいただき、ご支援いただきました。また、助言をいただいた諸先輩方、同回生の皆様にも感謝しています。最後に大学進学を支援してくれた両親に感謝の意を示し、本論文の結びと致します。

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] S. K. Baek, H. A. T. Kiet, and B. J. Kim, “Family name distributions: Master equation approach”, Phys. Rev. E **76** (2007), 046113 1–7.

[6] D. H. Zanette and S. C. Manrubia, “Vertical transmission of culture and the distribution of family names”, Physica A **295** (2001) 1-8.

[7] S. C. Manrubia and D. H. Zanette, “At the Boundary between Biological and Cultural Evolution: The Origin of Surname Distributions”, J. theor. Biol. **216** (2002) 461–477.

[8] C. Scapoli, E. Mamolinia, A. Carrieria, A. Rodriguez-Larralde and I. Barrai, “Surnames in Western Europe: A comparison of the subcontinental populations through isonymy”, Theoretical Population Biology **71** (2007) 37–48.

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

[10] 日本の姓の全国データベース, http://www.ipc.shizuoka.ac.jp/˜jjksiro/kensaku.html.

[11] 明治安田生命 生まれ年別名ランキング,
http://www.meijiyasuda.co.jp/profile/etc/ranking/year_men/.

[12] M.E.J. Newman, “Power laws, Pareto distributions and Zipf’s law”, Contemporary Physics **46** (2005), 323–351.

付録　ReaDのデータ取得方法

Read のデータはカテゴリ検索＞研究者＞研究分野別から検索、よりすべての分野を選択し一覧表示させた。ここでは 40 人のデータが一画面に表示されるが、検索結果約 16 万人のデータを手作業で取得するのは困難である。そのため UWSC というツールを使用し、以下のようなマクロ操作を行った。

ReaD の画面を開いたブラウザとメモ帳を起動しておく

1. ブラウザ上ですべてを選択 (Ctrl+A) コピー (Ctrl+C)
2. メモ帳に移動 (Alt+Tab)
3. メモ帳に貼り付け (Ctrl+V)
4. ブラウザに移動 (Alt+Tab)
5. 「次へ」を検索 (Ctrl+F、次へ)

この操作を繰り返すことでデータを得た。

また得られた txt データをもとに、

- grep で行初めの文字が [0-9] であるものを取り出した。
- Terapad で全角スペース、全角括弧、全角括弧閉じを半角カンマに置換した。
- さらに grep で, の数が 5 つの行だけを取り出した。

こうして得られたデータは [数字, 姓, 名, セイ, メイ,] となっており、これをもとに解析を行った。