日本統計学会誌
第35巻，第1号，2005年9月
55頁～70頁

日本人の名字の統計解析

千　田　　敏*，間　瀬　　茂*

Statistical Analysis of Japanese Surnames

Satoshi Chida* and Shigeru Mase**

この論文では，現在入手可能な最大の日本人の名字データを解析する．多出姓に対する“順位-サイズ関係”をZipf分布，稀少姓に対する“サイズ-頻度関係”をYule分布を用いた当てはめで検証する．更に，その結果を用い，現存する名字の総数の推定を試みる．また，名字の総数が将来どのように変化するかをシミュレーションにより予想する．

We study the “rank-size” and the “size-frequency” relations for the largest data of Japanese surnames covering almost 63% of Japanese households. Using the result of fitting of Yule (Zipf) distributions to rare surnames, the total number of present Japanese surnames is estimated to be about 105,000 kinds. Also we predict how many rare surnames will extinct in future by a simulation.

序．名字データ

日本人の名字データに関する調査は，長く民間研究家による名簿等からの調査が主なものであった．柳田国男は名字総数を約8万と見積もっていたという．その集大成ともいうべき丹羽基二（1996）には全部で291,531種類の苗字が収録されている．但し，これは読みの違いや，漢字表記の微妙な違いも，全て異なるとして数えたものである．電子計算機による事務処理の進展に伴い，生命保険会社の顧客データから名字の種類と数の集計結果が公開（第一生命広報部（1987）参照）されるようになった．更に，1990年代に入り，NTTの電話帳を電子化したCD-ROMが商品化されると，それから名字の種類と頻度を悉皆集計することが試みられた．この論文では，そうした集計結果の代表例二つを併用したデータを使用する．その一つは村山忠重（2003）で紹介されている順位が30,000位までの名字のサイズデータであり，今一つは須崎春夫（須崎春夫氏のウェッブページ参照）が電子電話帳と個人ウェッブサイトで収集・集計したデータの内，該当サイズが100件以下の名字の頻度データである．この二つを併用することにより，29,727,887件，名字の総数で99,466種類という基礎データを得た（サイズ300件までの頻度を付録Bに紹介する）．

注意：名字は苗字，姓，氏とも呼ばれる．歴史的起源からいえば，それぞれ意味が異なるらしい．現在でも法律用語としては「氏」が用いらる．

NTTの電話帳や，その電子版を名字データのソースとして利用する際，問題となる幾つかの点がある．何よりも，これは文字通りには全国の世帯母集団からの無作為抽出とはみなせな

* 東京工業大学大学院情報理工学研究科，〒152-8550 東京都目黒区大岡山2-12-1 W8-28

い．但し，特定の名字の所有者が，電話帳への記載を好む，もしくは拒否するという事情は，特に特殊な名字を除き考えにくいので，この点は大きな問題にはならないと思われる．その他注意すべき点を，名字研究家の森岡浩氏のウェッブページを参考にまとめると

- 電話帳記載の個人名は本名もしくはその正しい表記とはかぎらない．重複する可能性もある，
- 電話帳には漢字の読みが記載されておらず，掲載位置から推測するほかない，
- 名字と名前の間の区切りのないものが多くあり，両者を区別することが困難なことがある．
- 電子電話帳データは，NTT の電話帳を OCR ソフトで処理して作るため，良く似た字を誤判読する可能性がある．
- 携帯電話の普及，およびプライバシー保護意識の高まりから，電話帳への掲載件数が減少しており，正確な調査のためには，1990 年代前半あたりの電話帳の利用が好ましい．

電子電話帳による集計では，同じ漢字表記を持つ名前は，読み方が異なっても，同一（起源）とすることが普通のようである．例えば『東海林』は『しょうじ』とも『とうかいりん』とも読むが，同一とされている．これは，電話帳への記載順序を見ればある程度区別可能であるが，『山崎』（やまざき，やまさき）のように微妙な違いのものも多く，きりがないという（村山忠重（2003）参照）．一方，『阿部，安倍，阿倍，安陪，安部』などのように，読みが同じでも漢字表記や字体が異なるものは，別々に数えている．当然，特殊な漢字や異体字は原則使われない．多出姓のサイズを，その順位に対してプロットすると図1のようになる．また，希少姓の世帯サイズと頻度をプロットすると図2のようになる．

2. Zipf 分布と Yule 分布

有限もしくは無限のカテゴリからなる集団に対し，各カテゴリの順位とサイズの関係（順位-サイズ関係）を考える．様々な分野のデータに付いて，ある実数 $a>0$ が存在して

$$(\text{順位})^a \times \text{サイズ} = \text{一定}$$

という順位-サイズ関係が成り立つことが，これまで報告されてきた．たとえば，1冊の本に含まれる単語の数（レーマン（1984）参照），アメリカの都市の人口（レーマン（1984）参照），

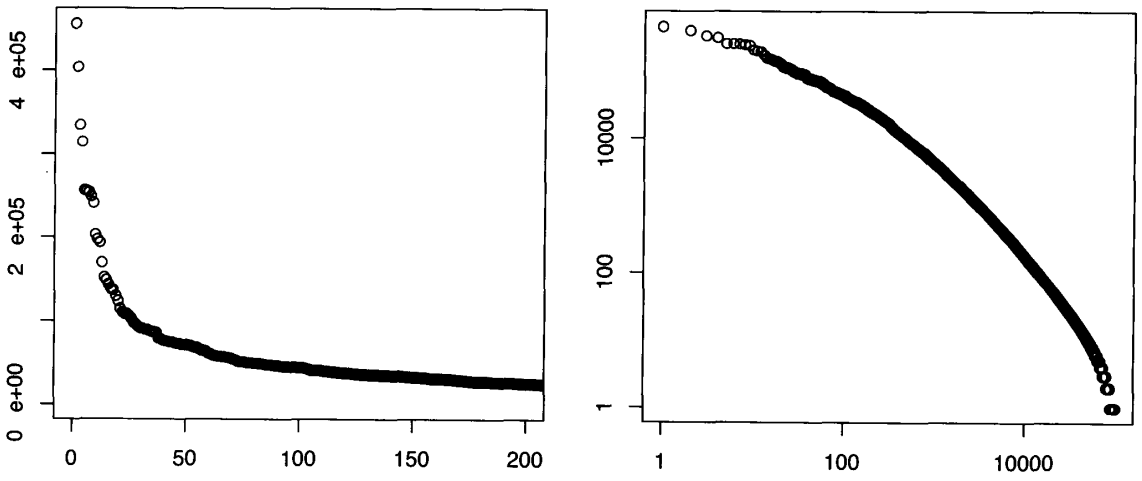

図1 多出姓の順位 x と世帯サイズ y のグラフ（左）と両対数グラフ（右）

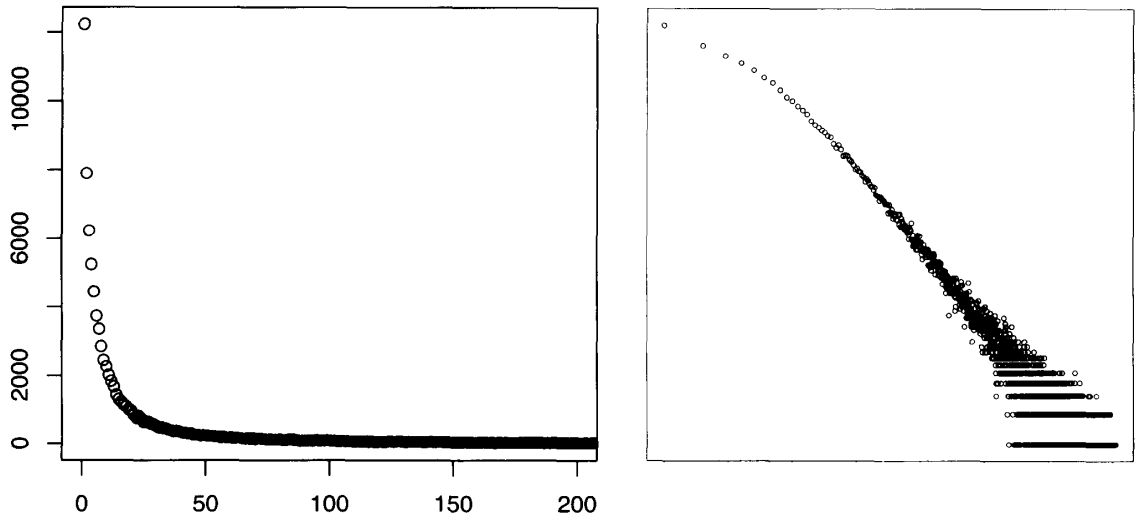

図2　希少姓の世帯サイズ x と頻度 y のグラフ（左）と両対数グラフ（右）

ウェッブページに貼られたリンクの数（Aiello, W. et al.（2002）参照）などで印象的な例がみられる．これを最初に提唱したアメリカの言語学者 G.K. Zipf に因み Zipf の法則と呼ぶ（Zipf（1949）参照）．

Zipf の法則は，順位 $x=1, 2, \ldots$ に対するカテゴリの出現サイズの確率が

$$f_Z(x; a)=C/x^a \tag{1}$$

となることを意味する．ここで正規化定数 C はツェータ関数 $\zeta(a)$ の逆数である．これを Zipf 分布と呼ぶ．この確率関数が全ての順位で意味を持つためには $a>1$ である必要があるが，Zipf が最初に提唱した形では $a=1$ であり，有界な順位範囲でしか意味を持たない．Zipf 分布の確率関数は両対数グラフで表現すれば，傾き $-a$ の直線 $\log y=-a\log x+\log C$ となる．アメリカ，中国，イギリスのマン島における名字の分布に対して Zipf 分布を当てはめた例がある（佐藤葉子，瀬野裕美（2003）参照）．それらによれば，Zipf 分布は部分的には良い当てはまりを示すものの，全体としての当てはまりは必ずしも良くない．

Mase（1992）は，名字データへのあてはまりを良くするために次の修正型 Zipf 分布を提案した．

$$f_{mZ}(x; a, b, c)=C\frac{c^x}{(x+b)^a},\quad C^{-1}=\sum_{x=1}^{\infty}\frac{c^x}{(x+b)^a}. \tag{2}$$

この確率分布は，$c<1$ ならば $0<a\leq 1$ であっても意味を持つ．とくに $c=1$ の場合は Zipf-Mandelbrot 分布と呼ばれることがある．Mase（1992）は，より稀な名字のサイズの推定を目的として，生命保険会社による日本の名字上位 200 位までのサイズデータ（第一生命広報部（1987）参照）に対して，修正型 Zipf 分布によるあてはめを行い，全範囲で良好な当てはめ結果を得た．

Zipf 分布の連続型は Pareto 分布と呼ばれ，企業の規模の分布等として経済学でしばしば登場する．Zipf 分布および Pareto 分布に関しては，様々な一般化が提案されており，渋谷政昭（2003）による総合報告に詳しい．しかしながら Zipf 分布が様々な分野のデータの近似分布として登場する理由に付いては，幾つかの理論があるものの，結局の所，経験的事実と述べておくのが適当と思われる．Zipf 分布を様々なデータにあてはめた研究によると，しばしば上位幾

つかのあてはまりが悪いことが報告されている.

Zipf 分布と双対的な分布が Yule 分布である. Zipf 分布がサイズの大きいカテゴリの順位の分布とすれば, 逆にサイズが小さい稀なカテゴリーの頻度（サイズ-頻度関係）に着目する. サイズが丁度 x であるカテゴリーの頻度確率として, 次の Yule 分布が用いられることがある.

$$f_Y(x; a)=x^{-1/a}-(x+1)^{-1/a}, \quad x=1, 2, \ldots \qquad (3)$$

パラメータ a は正でありさえすれば良い. $1/a$ の形のパラメトライゼーションは, Zipf 分布と Yule 分布の双対性（付録 A 参照）を考慮してのものである. Yule 分布も Zipf 分布と同様に, $a\simeq 1$ ならば, 両対数グラフに描くとほぼ直線状となる（図 3）.

注意：分布（1）と（3）は区別されずともに単に Zipf 分布と呼ばれたり, Zipf の第一法則, 第二法則と区別されることがある. 渋谷政昭（2003）では, 確率分布（3）は $a=1$ の時は単に Zipf 分布, 一般の a では Zipf-Mandelbrot 分布とされ, 逆に Yule 分布は次の形の確率分布のこととされている.

$$f(x)=\alpha(x-1)!/(\alpha+1)^{\bar{x}}.$$

（1）と（3）のどちらか, もしくは双方を Yule 分布と呼ぶ文献もあり, 混乱している. この論文では単に区別の便宜上, （1）を Zipf 分布, （3）を Yule 分布と呼ぶことにする. レーマン（1984）ではジェームス・ジョイス著『ユリシーズ』に登場する語彙に Yule 分布および Zipf 分布が良く当てはまることが紹介されている. こうした事情が言語学者等が Zipf, Yule 分布に関心を寄せてきた理由であるが, 全ての小説で良い当てはまりが得られるわけではないことも指摘されている.

今回解析する名字データ, また従来の研究においても, Zipf 分布が近似的に当てはまるデータには, 同時に Yule 分布も近似的に当てはまることが多いことが知られている. Zipf 分布と Yule 分布は多数のカテゴリからなる大規模集団に対する, それぞれ, 多出カテゴリと, 希少カテゴリに対する分布であり, 意味的にもなんらかの双対性が予想される. しかし, 実際には順位に多くのタイが出現したり, 全てのサイズ 1, 2, ... に対して対応するカテゴリが存在するわけではない. こうした事情から, 両者の関係を理論的に厳密に議論することは困難である.

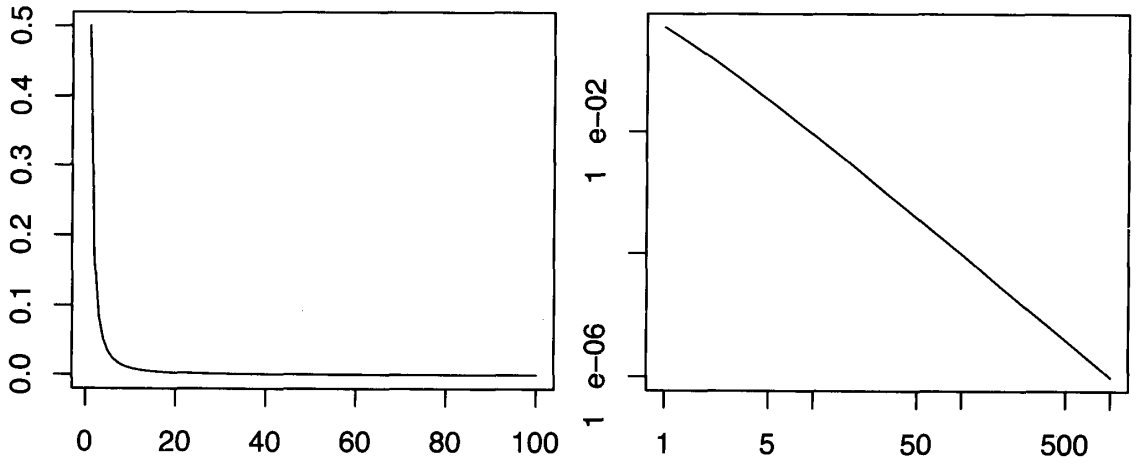

図 3 $a=1$ の Yule 分布の確率関数のグラフ（左）と両対数グラフ（右）

適当な文献も見当たらないようなので，参考のために，直感的なレベルで両分布の双対性を付録 A に示した．

3. 順位データへの Zipf 分布のあてはめ

順位で 1,000 位までの名字の世帯サイズ（電話帳記載件数）に，Zipf 分布（1）を最小自乗法を用いて当てはめてみる．すなわち，順位が i 位であるような世帯サイズ X_i の比率 x_i について，誤差の自乗和

$$S=\sum_{i=1}^{99,466}(x_i-f_Z(i;\,a))^2$$

を最小にする a を求める．最小自乗推定量は $\hat{a}=0.623$，決定係数は $R^2=0.965$ となった（図 4）．決定係数の値とグラフの様子から，よく当てはまっていると考えられる．しかし $\hat{a}<1$ であるため，（1）は密度関数とはならないことを注意する．

表 1 と図 5 は順位 1,000 位までと，それ以上の順位に別個に Zipf 分布を当てはめた結果である．従来の名字データの解析でしばしば注意されて来た Zipf 分布の広範囲での当てはまりの悪さを確認する結果となっている．

次に，全順位範囲に修正型 Zipf 分布（2）を当てはめてみる．修正型 Zipf 分布は $c<1$ ならば確率分布となることが保証されるため，最尤法を用いる．つまり，対数尤度

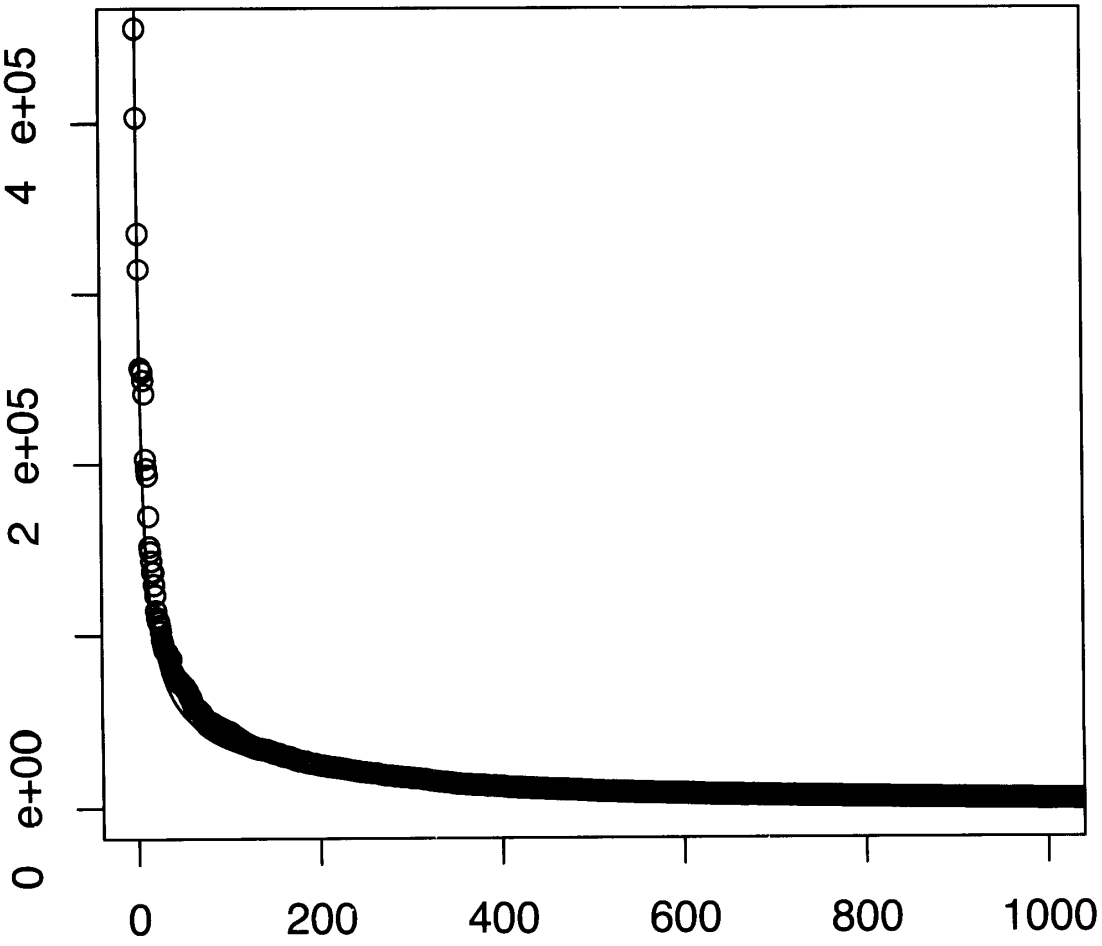

図 4　順位 1,000 位までの世帯サイズへの Zipf 分布のあてはめ

表1 Zipf分布のパラメータの最小自乗推定量と決定係数

	$\hat{a}$ の値	決定係数
順位1,000位まで	0.6229	0.9649
順位1,000位以降	1.450	0.9980

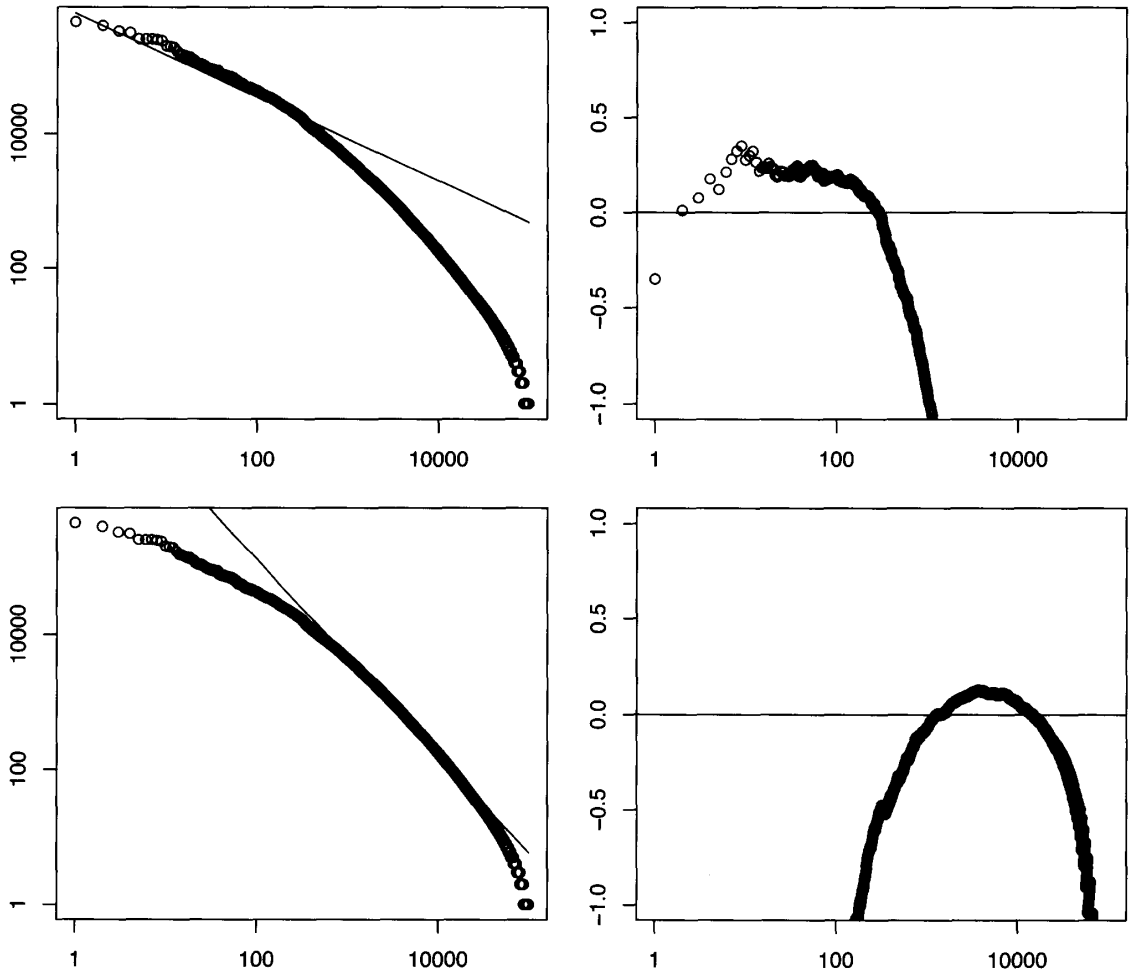

図5 順位1,000位まで（上）と，1,000位以降（下）へのZipf分布のあてはめ（左）とその相対誤差（右）

$$L(a, b, c) = \sum_{i=1}^{99,466} X_i \log f(i; a, b, c)$$

を最大にするパラメータを求める．統計解析システムRの汎用最適化関数optimを用いた最適化により，最尤推定値は

$$(\hat{a}, \hat{b}, \hat{c}) = (0.9789, 5.883, 0.9999)$$

となった．図6から分かるように，より広い範囲での当てはまりが確認できる．

4. 希少姓データへのYule分布のあてはめ

次に，希少姓のサイズと頻度のデータにYule分布（3）を当てはめてみる．1世帯から始まる希少姓にYule分布を当てはめた結果（図7上）から，おおよそ世帯サイズ20を境とし，それ以下では大きく乖離することが分かる．そこで，世帯サイズ20件以上1,000件以下の部

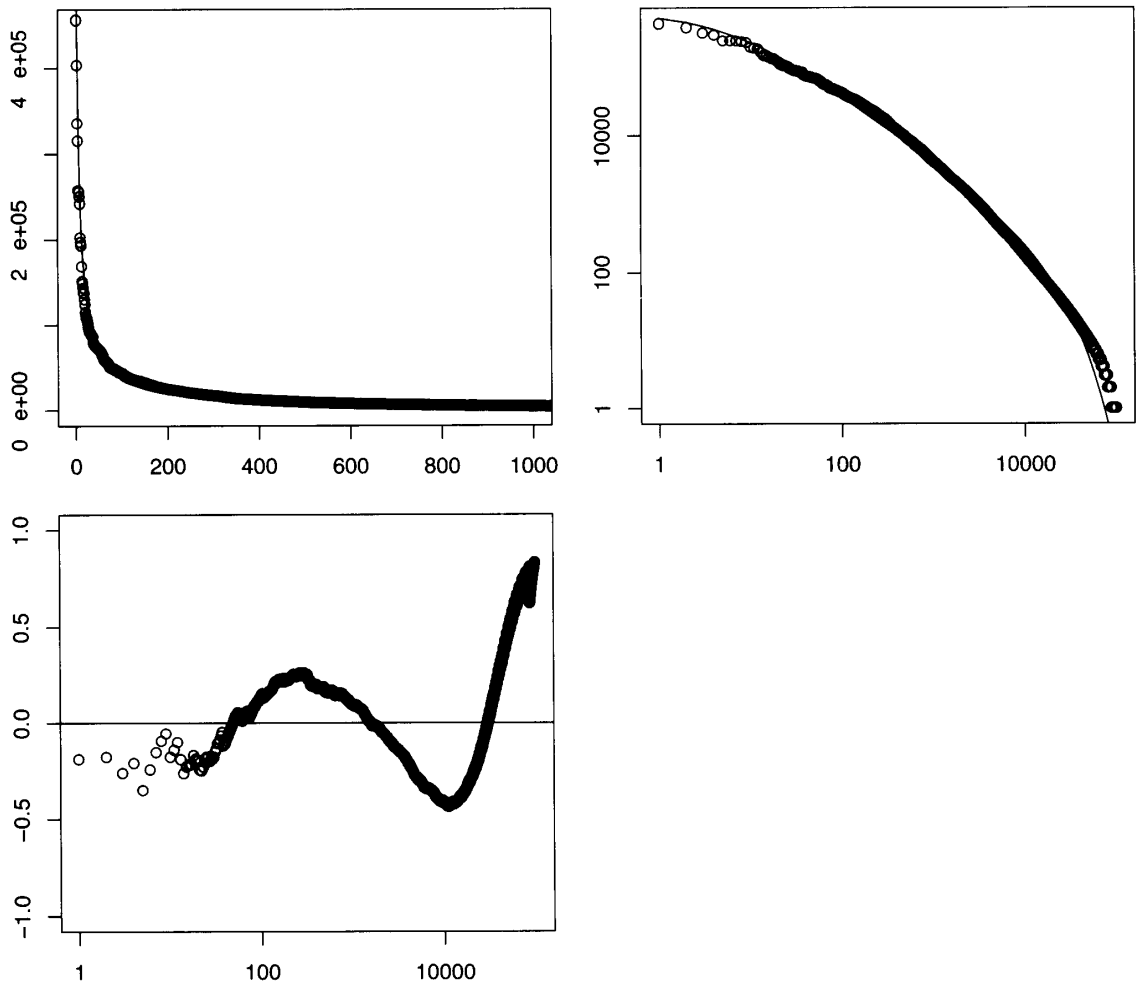

図 6　修正型 Zipf 分布の当てはめ（上左），両対数グラフ（上右），及び相対誤差（下）

分に条件付き Yule 分布

$$f_{Y:20}(x)=20^{1/a}(x^{-1/a}-(x+1)^{-1/a}),\quad x=20, 21, \ldots, 1{,}000 \qquad (4)$$

を最尤推定法で当てはめてみる．最尤推定量は $\hat{a}=1.633$ となった．図 7（下）より，20 世帯以上では比較的良い当てはまりが確認できる（参考のため 20 世帯未満まで延長して描いてある）．

5. 日本の名字の総数の推定

今回用いたデータの総数は 29,727,887 件である一方，2000 年度国勢調査（総務省統計局のウェッブページ参照）による日本の総世帯は 47,062,743 件であり，およそ 1,700 万件の世帯が電話帳に記載されていないことになる．従って，電話帳に 1 件しか記載されていない名字でも，実際には 2 世帯以上存在する可能性があり，電話帳に 1 件も記載されていない名字も存在する可能性がある．電話帳にちょうど i 件記載されている名字の頻度を Y_i，全国にちょうど i 世帯存在する名字の数を X_i とする．簡略化のために，各世帯が電話帳に電話番号を記載する確率 p は，世帯によらず一定であり，互いに独立であるとする．p の推定値として

$$\frac{\text{電話帳記載の総件数}}{\text{日本の総世帯数}}=\frac{29{,}727{,}887}{47{,}062{,}743}=0.632\cdots \qquad (5)$$

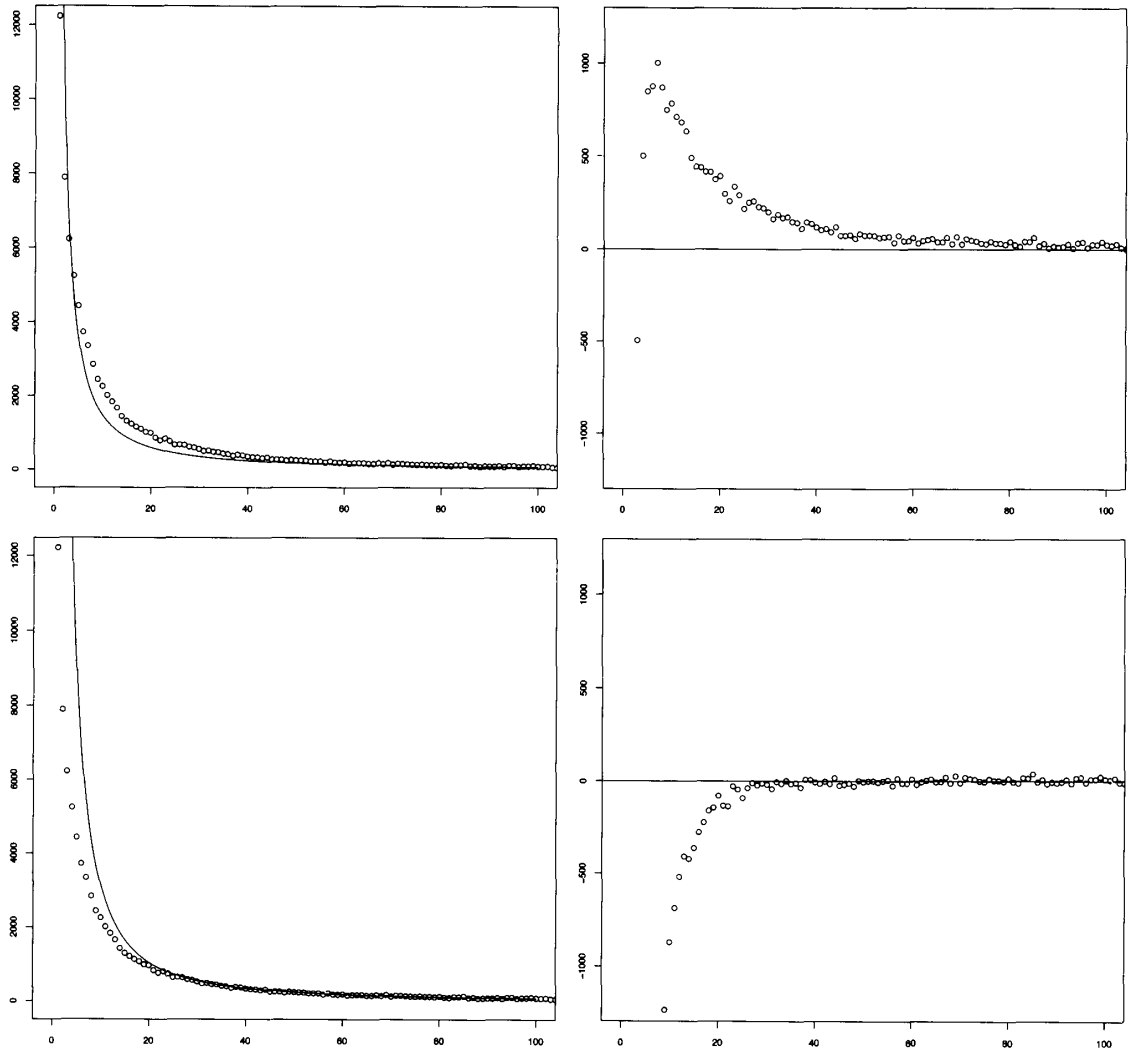

図7 希少姓へのYule分布の当てはめ．全範囲への当てはめ結果（上左）とその絶対誤差（上右）．20世帯以上への条件付きYule分布の当てはめ（下左）と絶対誤差（下右）

を用いる．日本全体でk世帯存在するある名字が，電話帳にちょうどi件記載される確率は，仮定の下で

$$\binom{k}{i}p^i(1-p)^{k-i}$$

となり，したがって関係

$$Y_i=\sum_{k=i}^{454,630}\binom{k}{i}p^i(1-p)^{k-i}X_k+\epsilon_i \qquad (6)$$

が得られる．ここでϵ_iは，実際のデータY_iとの食い違いを表す．454.630は最大姓佐藤の件数である．全てのiでY_iそしてX_iは正とは限らないが，少なくとも$i\leq 300$では$Y_i>0$で，ほぼ単調減少であることを注意しておく．

データ$Y=(Y_1, Y_2, \ldots, Y_n)^t$を用い，式（6）によって

$$X=(X_1, X_2, \ldots, X_n)^t$$

を推定することを試みると，回帰式 $Y=A_nX+\epsilon$ が得られる．ここで計画行列 $A_n=(a_{ij})$ は

$$A_n=\begin{pmatrix} p & {}_2C_1 p(1-p) & {}_3C_1 p(1-p)^2 & \cdots & {}_nC_1 p(1-p)^{n-1} \\ 0 & p^2 & {}_3C_2 p^2(1-p) & \cdots & {}_nC_2 p^2(1-p)^{n-2} \\ 0 & 0 & p^3 & \cdots & {}_nC_3 p^3(1-p)^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & p^n \end{pmatrix}$$

である．しかしながら，この線形重回帰式は説明変数と目的変数の数が同じであり，最小自乗推定値の精度は当然低くなる．更に A_n に $p=0.632\cdots$ を代入すると，例えば $Y_n=(0.632\cdots)^nX_n+\epsilon_n$ と X_n の係数が極めて小さくなり，X_n の推定値は一層不安定にならざるを得ない．この問題のように，不完全なデータから母集団の種類数や稀少種の分布を推定する問題は，個票の開示に関連する問題として研究されており，一般に解くのが困難であることが知られている（渋谷政昭（2003）参照）．先の議論から，少なくとも $Y_{21}, Y_{22}, \ldots, Y_{300}$ は条件付き Yule 分布に比較的良く適合することが分かった．$X_{21}, X_{22}, \ldots, X_{300}$，についても，同じパラメータの条件付き Yule 分布 $f(i)$ が当てはまると仮定することは，合理的と考えられる．回帰式

$$Y_i=a_{i1}X_1+a_{i2}X_2+\cdots+a_{i300}X_{300}+\epsilon_i$$

の変数 $X_{21}, \ldots, X_{300}$ を $X_{21}f(21)/f(21), \ldots, X_{21}f(300)/f(21)$ で置き換えると，回帰式

$$Y=A^*X+\epsilon, \quad X=(X_1, X_2, \ldots, X_{20}, X_{21})^t \tag{7}$$

を得る．ここで，計画行列 $A^*=(a^*_{ij})$ は成分

$$a^*_{ij}=a_{ij}, \quad 1\leq i\leq 300, 1\leq j\leq 20,$$
$$a^*_{i21}=\frac{1}{f(21)}\sum_{j=21}^{300}a_{ij}f(j), \quad 1\leq i\leq 300$$

を持つ 300×21 行列である．村山・須崎データを用いた回帰結果は表 2，および図 8 の様になった．自由度調整済み決定係数は $R_{\mathrm{adj}}=0.994$ となった．但し，制約付きの最小自乗法であるため，R_{adj} は制約下での最小誤差自乗和を用いて計算した．推定値と電話帳データとの差を見てみると，電話帳データ数が推定総世帯データ数を上回っている箇所がある．これは，例えば電話帳に 1 件しか記載されていない名字が，実際は 2 世帯以上ある名字に由来する可能性があることによる．

次に，こうして得られた稀少姓の総世帯数に関する推定値 X から，現存していながら電話帳に記載されていない名字の種類数 Y_0 を推定する事ができる．丁度 i 世帯存在する名字について，i 世帯全てが電話帳に記載しない確率は $(1-p)^i$ であるから

$$Y_0=\sum_{i\geq 1}(1-p)^iX_i \tag{8}$$

となる．（8）式に X の推定値を代入すると $Y_0=5{,}432$ となる．これにより，電話帳に記載されていない名字の種類数は 5,432 種類，現存する名字の種類数は 99,466＋5,432＝104,898 種類と推測される．仮に X データと Y データが同一としても，ほぼ同じ値 $Y_0=6023.0$ が得られることを注意しておく．

注意：線形重回帰式（7）を単純に最小自乗法で解くと，負の X_i 等を含む無意味な解しか

表2 50位までの日本の稀少姓世帯数推定値（X_i）と電話帳記載件数（Y_i）

サイズi	X_i	Y_i	i	X_i	Y_i	i	X_i	Y_i
1	10823.0	12219	18	1310.4	1092	35	453.2	423
2	7766.5	7890	19	1173.3	1006	36	433.4	408
3	5180.2	6232	20	1020.6	981	37	414.9	366
4	4878.8	5243	21	1017.6	848	38	397.6	392
5	4222.6	4440	22	945.6	775	39	381.5	376
6	3876.1	3728	23	881.6	823	40	366.5	349
7	3871.0	3344	24	824.3	750	41	352.3	327
8	2969.5	2840	25	772.8	653	42	339.1	326
9	2606.6	2440	26	726.3	663	43	326.6	301
10	2567.3	2257	27	684.2	650	44	314.8	322
11	2039.6	2012	28	645.9	601	45	303.7	269
12	2037.5	1841	29	610.9	578	46	293.3	264
13	2010.7	1676	30	578.9	541	47	283.4	261
14	1723.2	1436	31	549.6	488	48	274.0	236
15	1565.2	1306	32	522.6	500	49	265.2	258
16	1476.0	1231	33	497.7	467	50	256.7	244
17	1313.6	1147	34	474.6	461			

得られない．意味のある解を得るために，拘束条件

$$X_i \geq X_{i+1}, \quad 1 \leq i \leq 20,$$
$$0.8\,Y_i \leq X_i \leq 1.2\,Y_i, \quad 1 \leq i \leq 21$$

を課し，Rの制約付き最適化関数constrOptimで解を求めた．但し，これだけでは，推定されたX_i, $i \leq 21$，とそれから求めたX_i, $i > 21$，の間のギャップが依然相当大きくなるため，両者がほぼ同じになるように調整した．

6. 名字の継承と断絶

名字は，親から子供に代々継承されるという意味で，正しく文化・社会的遺伝子（meme）の代表といえる．生物と同様，継承する子供がいなければ，名字は断絶する可能性が常にある．須崎データに登場する，全国でも数世帯という希少姓はとくに数世代で断絶する可能性が大きいといえよう．この節では，Galton-Watson型分枝過程モデルによるシミュレーションにより，将来の日本における名字の種類数の変化を予測してみたい．

名字の継承という異色の視点からGalton-Watson型分枝過程モデルを詳しく解説した文献に，佐藤葉子，瀬野裕美（2003）がある．Galton-Watson型分枝過程では，一つの世帯を出発点とし，世代交替毎に名前を継承する次世代の世帯数がランダムに増減すると考える．更に，次の強い仮定

- 時間は世代単位で数え，同世代の世帯は一斉に次世代世帯をもうけ，そして死亡する，
- 次世代の世帯数の増大・減少は単純マルコフ過程に従い，推移確率は全ての世帯で同一

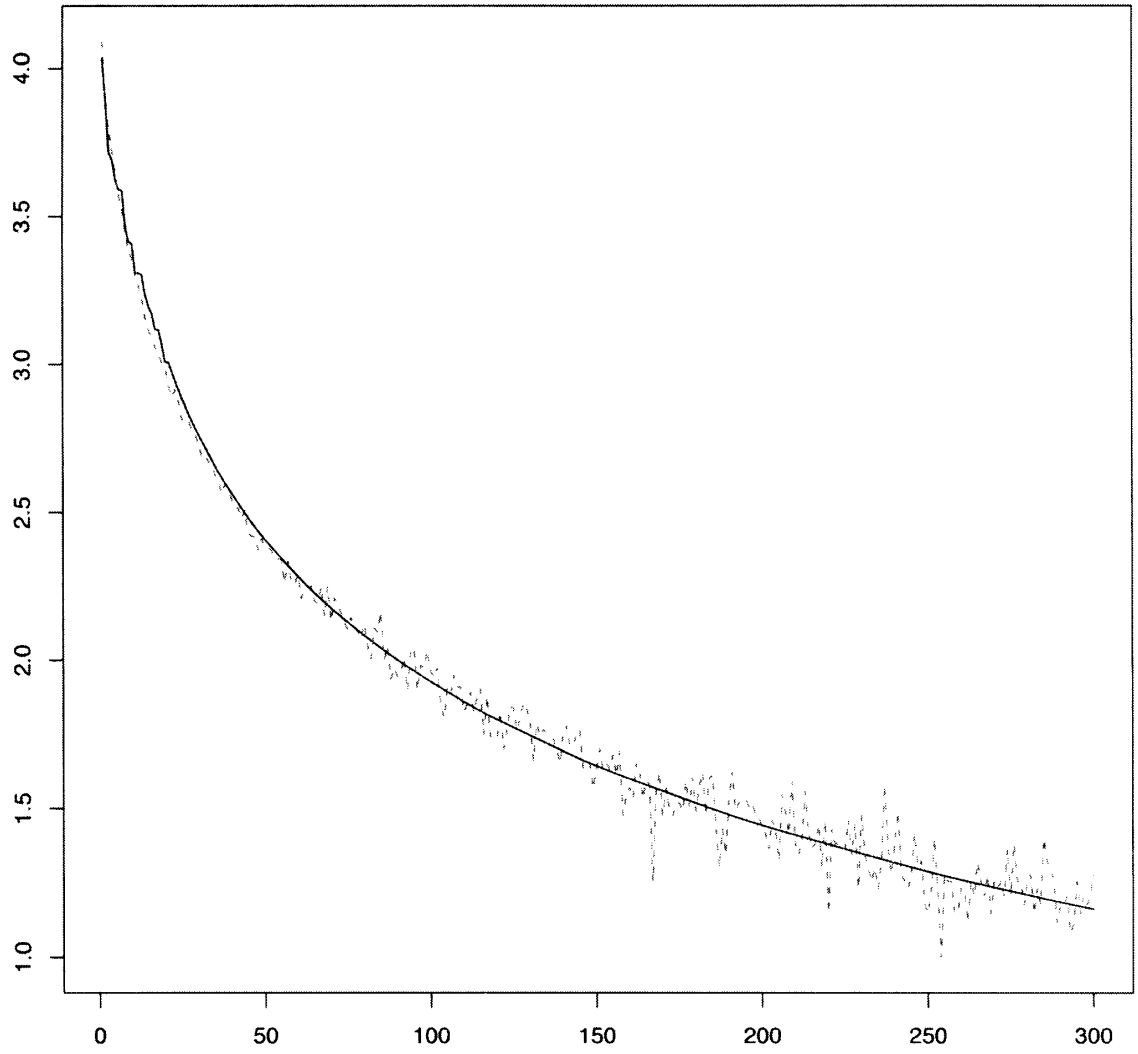

図 8　電話帳未掲載を含む稀少姓世帯頻度の推定結果．点線は掲載姓データ．横軸は世帯数，縦軸は該当名字数の常用対数値

を置く．

シミュレーションにあたっては，Galton-Watson 型分枝過程のパラメータ p_k, $k=0, 1, \ldots$ が必要になる．p_k は一つの世帯が，その名字を継承する k 世帯の次世代を持つ確率である．以下では，佐藤葉子，瀬野裕美（2003）で紹介されている成人男性 10,000 人あたりの出生男子数の分布を p_k とした（表 3 参照）．この出生男子数分布は，結婚平均持続期間が 15〜19 年の世帯の出生児数に，出生性比，未婚率を考慮にいれ，1992 年に厚生省により算出されたものである．

この $\{p_k\}$ と，先に求めた全国の稀少姓世帯数推定値を初期値とし，シミュレーションを行った．初期世代の各世帯に関し，世代交替ごとにパラメータ $\{p_k\}$ による増減を繰り返し，5 世代までの世帯数と名字分布を調べた．結果は表 4 のようになった．総数 20 世帯までの名字について，5 世代後までの予測世帯数を掲載している．1 世代後に 5 千種あまりの名字が消滅するという結果が注目される．希少姓は存続しにくいという結果は，図 9 から一層はっきりする．

7. 考　察

現在入手可能な最大の日本人の名字データを基に，多出姓の世帯サイズへの Zipf 分布の当

てはめ，希少姓の頻度へのYule分布の当てはめを行った．Zipf分布に関しては，従来の同種の研究と同様に，上位（1000位程度）まででは一定の当てはまりが確認されたものの，より広い範囲では無理があることが確認された．一方，修正型Zipf分布は，一層広い範囲で良い当てはまりが確認された．希少姓へのYule分布当てはめは，小集団の悉皆的調査を除けば，そうしたデータそのものがこれまで得られにくかったことから，比較すべき研究は無いようである．今回の調査でも，広範囲での良い当てはまりは確認されなかったが，世帯サイズ20件以上に限れば，ある程度の当てはまりを確認できた．

なぜ名字データにZipf，Yule分布が当てはまるのかは，名字の起源と継承の多様さを考えれば，とりあえず経験的事実といっておくよりしかたがないと思われる．歴史的に，そして明治の新姓採用時においても，既に世帯数が多かった名字程，一層多くの人が自分の名字として採用することが多かったであろうという，容易に想像される背景がヒントになるかも知れない．一方，希少姓データへの部分的なYule分布当てはめについては，Zipf分布とYule分布の間の双対性がヒントになる．その上で，なぜ20位以下の希少姓では当てはまりが悪いかについては，おそらく明治新姓採用時に恣意的に作られた多数の非伝統的な名字の一斉の出現と，それ以降実質的に新しい名字の誕生が絶たれたこと，そして明治以来数世代のGalton-Watson過程的な経過では定常的な安定分布への推移がまだ見られない，等の理由が考えられるであろう．従来，名字に関する研究は学問的な対象とされることが少なく，主として民間研究家の個人的な努力に委ねられてきた．その理由は，日本人の相当部分を網羅するようなデータが比較的最近まで存在しなかったこと，そして日本人の名字の相当部分（9割以上ともいわれる）が，明治始めに恣意的に選ばれた（1870年の「平民苗字許可令」で平民も名字を名乗ることが許され，更に1875年の太政官布告「平民苗字必称義務令」で名字が義務化された）という広く流布している意見が，背景にある．漢字表記とその読みの多様性が名字の単位の特定を難しくし

表3　実データに基づいた次世代の名字継承世帯数の分布 p_k

k	0	1	2	3	4	5	6
p_k	0.32840	0.38670	0.23760	0.04327	0.00357	0.00030	0.00003

表4　稀少姓数の世代別変化のシミュレーション結果．5世代後までの世帯数20以下の名字の総数

	消滅姓累積数	1	2	3	4	5	6	7	8	9	10
現世代	0	10823	7766	5180	4878	4222	3876	3871	2969	2606	2567
1世代後	4644	7275	7186	5533	4606	4127	3708	3380	3054	2561	2398
2世代後	8100	5750	6225	5283	4522	3939	3420	3150	2886	2601	2340
3世代後	10881	4966	5468	4735	4299	3777	3423	3070	2730	2553	2263
4世代後	13357	4264	4819	4387	4054	3475	3233	3005	2711	2399	2246
5世代後	15527	3722	4394	3999	3694	3342	3218	2883	2606	2373	2199
		11	12	13	14	15	16	17	18	19	20
現世代		2039	2037	2010	1723	1565	1476	1313	1310	1173	1020
1世代後		2154	2021	1859	1688	1546	1481	1383	1264	1217	1040
2世代後		2112	2051	1839	1689	1488	1434	1415	1247	1168	1144
3世代後		2054	1949	1755	1646	1601	1463	1354	1268	1094	1116
4世代後		2101	1928	1761	1581	1565	1443	1353	1247	1083	1111
5世代後		2048	1860	1806	1611	1481	1329	1311	1212	1118	1035

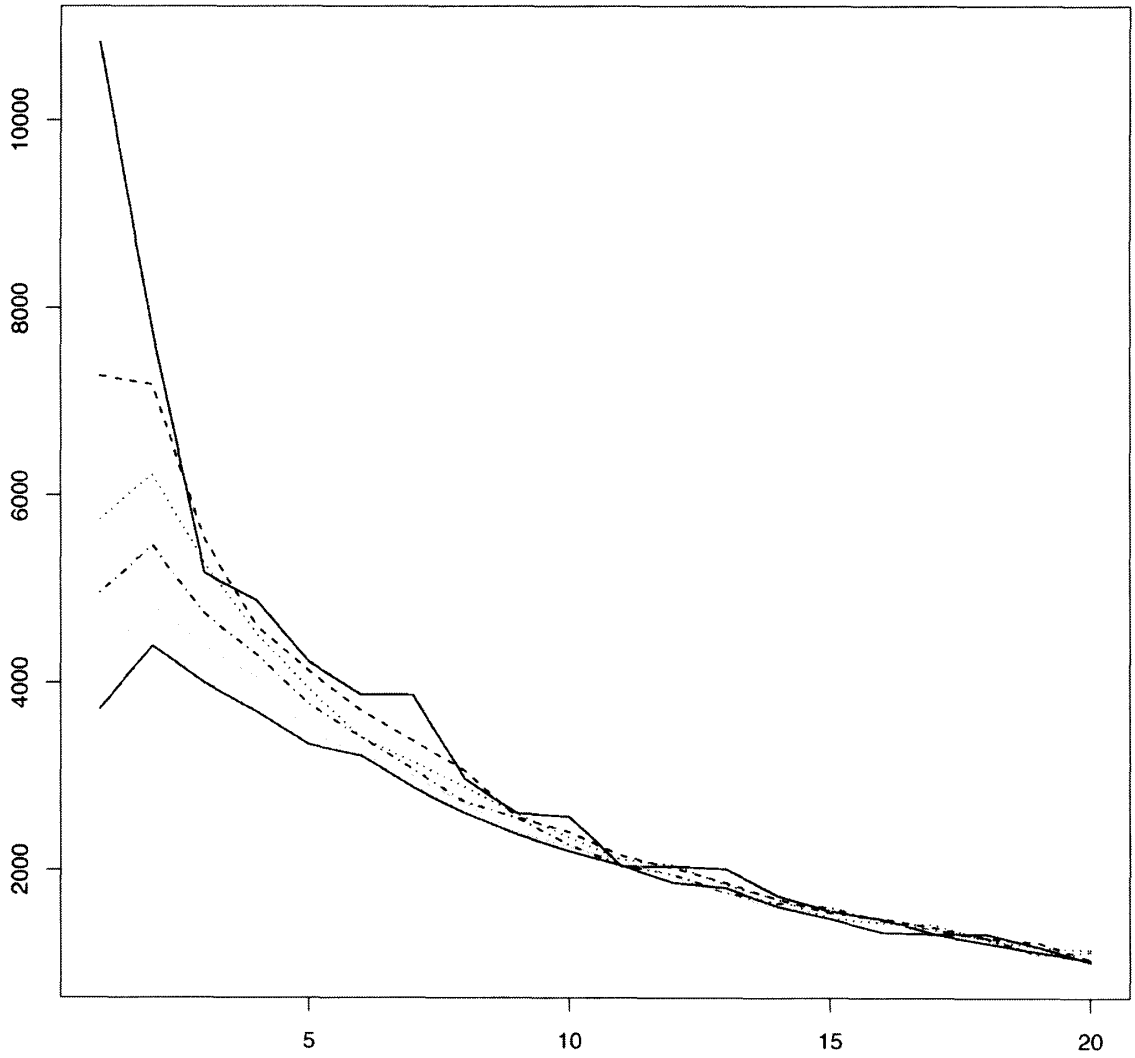

図9　稀少姓世帯数の世代毎のシミュレーション結果．一番上の実線グラフは現在の稀少姓世帯数推定値，以下世代交替毎にグラフはほぼ下降する

ていること，読みや表記の変更も含め，歴史的に日本人が名字を簡単に変更してきた，という事実もあげられる．国勢調査結果や，住民票の集計による名字分布データが得られる見込みが現状では無い以上，電子電話帳掲載のデータの悉皆調査が可能な最大のソースである事情は今後も変わらないであろう．一方で，携帯電話の急速な普及や，プライバシー意識から電話帳に電話番号を記載しない人の数は今後もますます増えると思われる．したがって，今回の調査で利用した，携帯電話の本格的普及直前の電子電話帳の集計結果が，名字に関する最も重要なデータといえるであろう．

最後に，今回の研究結果を踏まえ，幾つかの結論を私見としてまとめておきたい．Yule 分布の当てはめが全国世帯数で 20 件以下の名字で失敗するという結果は，これらこそが真の希少姓であり，おそらくそのかなりのものが明治の始めに多かれ少なかれ恣意的に全く新規に造られたか，歴史的な名字を改変して造られた名字であることを示唆すると思われる．一方，多出姓の相当広範囲な部分で（修正）Zipf 分布が良い当てはまりを示すという結果は，明治新姓の成立が，しばしば信じられているような全くの恣意的なものであったのではなく，その相当数がなんらかの組織的な由来を持つことを示唆するように思われる．公には名字を持たないはずの多くの庶民が，実際には名字を私称していた（武光誠（1998）参照），もしくは地域社会

にゆかりのある伝統姓を組織的に名乗った，等の背景が考えられる．

また，従来の名字総数の見積りの大幅な違いに付いても，今回のシミュレーション結果が示唆的と思われる．従来の調査の非系統的な性格や，名字単位の曖昧さを除いても10万から30万という予想の幅は大きすぎると思われる．今回のシミュレーション結果が示す，今後1世代で5,000種程，5世代では16,000種程の名字が消滅するという見積りは，逆に，明治8年当時から現在に至るおおよそ3～5世代の経過のうちで失われた希少姓の数が，相当なものであったことを強く示唆する．シミュレーション結果を単純に過去に3次スプライン補間すれば，例えば過去1世代および2世代の間にそれぞれ6,300種類，15,000種類あまりの名字が失われたという結果を示す．もちろん，世帯毎の名字継承数分布は，この一世紀あまりの間に劇的に変化していることや，大戦中の死亡数を考慮すれば，この数字は単なる参考にすぎないが，それでも万単位の名字がこの一世紀の間に失われたことを示すように思われる．従来の名字総数の見積りの大幅な食い違いの原因の一つは，こうした名字数のダイナミックな変化であろう．また，名字研究家の森岡浩氏によると，文献に珍姓として記載されている名字のなかに，電話帳にまったく記載例がない，幽霊名字ともいえるものが数多くあるという（森岡浩氏のウェッブページ参照）．小説等に登場した架空の名字を実在すると混同したものもあるようだが，こうした幽霊名字は，今回推定された5千種あまりの電話帳未記載名字であるか，かって実在したものの，既に消滅した名字が相当含まれていると考えてよいであろう．

謝辞．今回の研究は，多くの名字研究家のこれまでの地道な調査があって始めて可能になった．特に，須崎春夫氏からは，最も集計が困難な希少姓の膨大な調査結果を，著者の求めに応じ快く提供して頂いただいたばかりか，更にそのデータを公開する許可も頂戴しました．深く感謝します．

参考文献

[1] Abello, W. et al. (2002). Random evolution in massive graphs, in *Handbook of Massive Data Sets* (J. M. Abello et al., eds.), 97-122, Kluwer Academic Pub.
[2] 佐藤葉子，瀬野裕美（2003）．姓の継承と絶滅の数理生態学，京都大学学術出版会．
[3] 渋谷政昭（2003）．孤立個体数の推測，統計数理，**51-2**, 261-295.
[4] 須崎春夫氏のウェッブページ，URL http://www2s.biglobe.ne.jp/~suzakihp/index.htm.
[5] 総務省統計局の国勢調査に関するウェッブページ，URL http://www.stat.go.jp/data/kokusei/.
[6] 第一生命広報部編（1987）．日本全国苗字と名前おもしろBOOK，恒友出版．
[7] 武光誠（1998）．名字と日本人 先祖からのメッセージ，文芸春秋．
[8] 丹羽基二（1996）．日本苗字大辞典，芳文館出版部．
[9] Mase, S. (1992). Approximations to the Birthday Problem with unequal occurrence probabilities and their application to the surname problem in Japan, *Annals of Inst. Statist. Math.*, **44-3**, 479-499.
[10] 森岡浩氏のウェッブページ，URL http://home.r01.itscom.net/morioka/myoji/.
[11] 村山忠重（2003）．日本の苗字ベスト30000，新人物往来社．
[12] 村山忠重氏のウェッブページ，URL http://www.climbcom.com/m/.
[13] レーマンE. L. 監修，安藤洋美監訳（1984）．統計学講話，現代数学社．J. M. Tanur (1989). *Statistics: A Guide to the Unknown* (3rd ed.), Brooks/Cole Pub. Co.
[14] R Development Core Team (2005). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.
[15] Zipf G. K. (1949). *Human Behaviour and the Principle of Least-Effort*, Addison-Wesley, Cambridge MA.

付録 A. Zipf 分布と Yule 分布の双対性

Zipf 分布と Yule 分布は多数のカテゴリからなる大規模集団に対する，それぞれ，瀕出するカテゴリと，稀なカテゴリに対する分布であり，意味的にもなんらかの双対性が予想される．しかし，実際には順位に多くのタイが出現したり，全てのサイズ 1, 2, ... に対して対応するカテゴリが存在するわけではない．こうした事情から，両者の関係を理論的に厳密に議論することは困難である．適当な文献も存在しないようなので，参考のために，直感的なレベルの議論で Zipf 分布と Yule 分布の双対を以下に示しておきたい．

まず，$f(i)=C/i^a$, $i=1, 2, ...$, をパラメータ $a>1$ の Zipf 分布とする．十分大きなデータ数 N に付いて順位 i のカテゴリのサイズはほぼ NC/i^a となる．従って，サイズがそれぞれ $n+1, n$ である順位をそれぞれ j, k とすれば，近似的関係 $n+1\simeq NC/j^a$, $n\simeq NC/k^a$ がそれぞれ成り立つ．つまり，$j\simeq(NC)^{1/a}/(n+1)^{1/a}$, $k\simeq(NC)^{1/a}/n^{1/a}$ となる．これよりサイズが n であるカテゴリの頻度は $k-j\simeq(NC)^{1/a}/(n^{-1/a}-(n+1)^{-1/a})$ と見積もられる．これは Yule 分布に他ならない．

逆に，カテゴリの順位の分布が単調減少な連続関数 $f(x)$, $x>0$, を用いて，$f(i)$, $i=1, 2, ...$, と表されていると仮定する．直前と同じ状況を考えると近似的関係 $j\simeq f^{-1}((n-1)/N)$, $k\simeq f^{-1}(n/N)$ が成り立つ．実際には順位にはタイが生じるため，より正確にいえば $j(k)$ はサイズが $n+1(n)$ であるような順位の最大値と考えていることに注意しよう．従って，サイズが n であるカテゴリのサイズ $k-j$ はほぼ $f^{-1}(n/N)-f^{-1}((n+1)/N)$ となる．もしこれが Yule 分布に従うと仮定すれば，結局ある共通定数 D があり，近似的関係

$$f^{-1}(n/N)-f^{-1}((n+1)/N)\simeq D(n^{-1/a}-(n+1)^{-1/a}),\quad n=1, 2, ...$$

が成り立つことになる．従って，更に

$$\begin{aligned} f^{-1}(n/N)&=\sum_{i\geq n}(f^{-1}(i/N)-f^{-1}((i+1)/N))\\ &\simeq\sum_{i\geq n}(Di^{-1/a}-D(i+1)^{-1/a})\\ &\simeq Dn^{-1/a}\end{aligned}$$

が導かれる．つまり，$x=Dn^{-1/a}$ の形の実数に対して

$$f(x)\simeq\frac{D^a}{N}x^{-a}$$

となる．これは Zipf 分布に他ならない．

付録 B. 希少姓データ

一万位までの名字の世帯サイズデータは村山忠重氏のウェッブページにある．表 5 は希少姓データである．1 件から 100 件までは須崎春夫氏の調査による．101 件以上は村山忠重 (2003) から編集した．須崎氏のデータは，電子電話帳データを基本に，個人的に収集し実在を確認した電話帳未記載姓を加えたものである．

表5 稀少姓の頻度データ．世帯サイズiと頻度Y_i

i	Y_i	i	Y_i	i	Y_i	i	Y_i	i	Y_i	i	Y_i
1	12219	51	240	101	87	151	50	201	26	251	17
2	7890	52	234	102	94	152	43	202	23	252	25
3	6232	53	219	103	71	153	45	203	29	253	16
4	5243	54	219	104	64	154	43	204	27	254	10
5	4440	55	218	105	80	155	48	205	21	255	19
6	3728	56	182	106	76	156	41	206	35	256	18
7	3344	57	216	107	89	157	49	207	31	257	18
8	2840	58	185	108	81	158	30	208	26	258	14
9	2440	59	180	109	81	159	35	209	39	259	14
10	2257	60	195	110	67	160	37	210	24	260	17
11	2012	61	162	111	71	161	34	211	22	261	17
12	1841	62	172	112	78	162	45	212	25	262	13
13	1676	63	175	113	67	163	43	213	36	263	18
14	1436	64	178	114	75	164	34	214	26	264	16
15	1306	65	158	115	80	165	39	215	23	265	21
16	1231	66	156	116	56	166	38	216	24	266	18
17	1147	67	177	117	76	167	18	217	28	267	16
18	1092	68	141	118	55	168	36	218	20	268	19
19	1006	69	177	119	55	169	42	219	29	269	14
20	981	70	135	120	54	170	30	220	14	270	18
21	848	71	162	121	67	171	37	221	27	271	17
22	775	72	152	122	50	172	31	222	24	272	18
23	823	73	145	123	58	173	30	223	24	273	16
24	750	74	133	124	70	174	33	224	23	274	23
25	653	75	127	125	69	175	33	225	23	275	16
26	663	76	139	126	58	176	31	226	29	276	24
27	650	77	128	127	64	177	39	227	22	277	18
28	601	78	126	128	72	178	32	228	28	278	18
29	578	79	118	129	69	179	40	229	17	279	15
30	541	80	132	130	61	180	31	230	30	280	18
31	488	81	112	131	46	181	37	231	22	281	19
32	500	82	101	132	61	182	42	232	20	282	14
33	467	83	128	133	55	183	30	233	18	283	17
34	461	84	124	134	58	184	40	234	20	284	15
35	423	85	146	135	57	185	41	235	17	285	25
36	408	86	100	136	56	186	28	236	21	286	21
37	366	87	111	137	54	187	20	237	38	287	20
38	392	88	86	138	50	188	27	238	23	288	18
39	376	89	94	139	45	189	22	239	19	289	13
40	349	90	88	140	57	190	34	240	24	290	15
41	327	91	90	141	60	191	42	241	31	291	15
42	326	92	102	142	51	192	31	242	19	292	16
43	301	93	79	143	51	193	33	243	18	293	12
44	322	94	107	144	56	194	34	244	17	294	13
45	269	95	110	145	58	195	34	245	19	295	18
46	264	96	79	146	43	196	32	246	26	296	14
47	261	97	96	147	44	197	33	247	20	297	16
48	236	98	94	148	45	198	30	248	21	298	15
49	258	99	108	149	37	199	30	249	15	299	16
50	244	100	92	150	43	200	27	250	14	300	19