数理解析研究所講究録
第777巻 1992年 103-125

同姓問題 – 不等生起確率の誕生日問題

(Surname Problem – Birthday Problem with unequal occurence probabilities)

間瀬茂

(Shigeru MASE)

広島大学、総合科学部

(Fac. Integrated Arts and Sciences,

Hiroshima University)

1 序

比較的小人数の集団に同姓のメンバが偶々居ることを日常良く経験する。我々はこの事実をさして奇異とも思わない。このことはいわゆる「誕生日問題」、ある集団に同じ誕生日のメンバが居る確率、が我々に与えるパラドックスとさえ思える驚きと比べると対照的である。しかしながら、もし我々が日本における姓の異常な多さ (推計約 12 万種) と、各姓の占める比較的小さな割合 (最初の 5 千種併せても 92.3%) を考えると、「同姓問題」 (講演者の造語) は誕生日問題以上に不思議な事実なのである。おそらく構成メンバの姓は容易に知られるのに対し、誕生日は普通公表される事が無いという事情が、二つの問題に対する我々の受け止め方の差を生み出すのであろう。

以上のような背景を別にしても、n 人の集団に同姓のメンバが少なくとも一組居る確率を知ることは興味のある問題と思われる。容易に分かるように同姓問題は誕生日問題と深い関連があり、実際“不等生起確率の誕生日問題”と言い替えることが出来る。より一般的に同姓問題を次のように定式化出来る。$X_1, X_2, \ldots, X_n$ を同じ (有限もしくは無限) 離散分布 $P(X =$

$j) = p_j$, $j = 1, 2, \ldots$, を持つ独立な確率変数列とする。この確率変数中少なくとも一組同じ値を持つものが存在する 一致確率 R_n、全部が異なった値を持つ 不一致確率 $r_n = 1 - R_n$ を定義する。更に補助量として確率の巾和 $P_m = \sum_{i \geq 1} p_i^m$ を導入する。

同姓確率を具体的に求めるためには

1. (不) 一致確率の計算可能な理論式、もしくは近似式、

2. 日本の姓の頻度の統計

が必要になる。この講演では不一致確率 r_n の巾和 $\{P_m\}$ のベル多項式による表現と、ベル多項式の対数の形式的展開に基づくその近似式を紹介する。日本における姓の頻度に関するデータとしては第一生命がその顧客名簿 (832 万人分) を集計して得られた数値を用いる。但し公表されている頻度は上位 200 位迄のみであるため、これに一般化された Zipf 分布を当てはめることにより 201 位以上の姓の頻度を推定することにした。

我々の経験を裏付けるように、同姓確率 R_n は比較的小さな n でも相当大きな値をとることが分かる。例えば同姓確率が 50% を始めて越えるのは $n = 27$ (誕生日問題では $n = 23$)、90% を始めて越えるのは $n = 50$ (誕生日問題では $n = 41$) である。

今一つの応用として「真の誕生日問題」を考える。実際の誕生日の分布は一様とはほど遠く、倍程度くい違うことがある。1900 年から 1987 年に生まれた日本人の 1988 年における日別誕生日の分布を公表されたデータから推定した。実際推定された分布は最大・最少で 1.5 倍ほどの差がある。しかしながら、それから計算された誕生日確率そのものは、等確率の仮定から計算された誕生日確率とほとんど差が無かった。

2 ベル多項式

以下の議論ではベル多項式が主要な道具になるため、最初にその性質をまとめておく。詳しくは Comtet (1974) や Roman (1984) を参照されたい。. 以下の議論ではベル多項式が主要な道具になるため、最初にその性質をまとめておく。(指数型) 部分ベル多項式 $B_{n,k}(x_1, \ldots, x_{n-k+1})$ とは次の二重級数展開で定義される変数 $x_1, x_2, \ldots$ の多項式である;

$$\exp\left(u \sum_{m \geq 1} x_m \frac{t^m}{m!}\right) = 1 + \sum_{n \geq 1} \left\{ \sum_{k=1}^{n} u^k B_{n,k}(x_1, x_2, \ldots) \right\} \frac{t^n}{n!} \tag{1}$$

(指数型完全) ベル多項式 $Y_n(x_1, \ldots, x_n)$ は次の母関数で定義される;

$$\exp\left(\sum_{m \geq 1} x_m \frac{t^m}{m!}\right) = 1 + \sum_{n \geq 1} Y_n(x_1, x_2, \ldots, x_n) \frac{t^n}{n!}$$

つまり $Y_n = \sum_{k=1}^n B_{n,k}$ and $Y_0 = 1$ である。$B_{n,k}$ の具体的表現は次の様になる;

$$B_{n,k} = \sum \frac{n!}{a_1! \cdots a_n!} \prod_{i=1}^n \left(\frac{x_i}{i!}\right)^{a_i} \tag{2}$$

ここで右辺の和は $\sum_{i=1}^n a_i = k$ 且つ $\sum_{i=1}^n i\, a_i = n$ であるすべての $(a_1, a_2, \ldots, a_n)$ についてとる。特に $B_{n,n} = x_1^n$ 。次の性質は容易に分かる;

$$\begin{aligned} B_{n,k}(abx_1, \ldots, ab^j x_j, \ldots) &= a^k b^n B_{n,k}(x_1, \ldots, x_j, \ldots), \\ Y_n(bx_1, b^2 x_2, \ldots, b^n x_n) &= b^n Y_n(x_1, \ldots, x_n) \end{aligned} \tag{3}$$

ベル多項式は次の漸化式を持つ (Roman, 1984, Chap. 4.1.8);

$$Y_n(x_1, \ldots, x_n) = \sum_{i=1}^n \binom{n-1}{i-1} x_i Y_{n-i}(x_1, \ldots, x_{n-i}) \tag{4}$$

更に我々は多重添字のベル多項式を必要とする。この概念が既に文献中にあるかどうかは不明であるが、定義そのものは通常のベル多項式のそれの単純な拡張である。S_N を多重添字 $a = (a_1, \ldots, a_N) \neq (0, \ldots, 0)$ の集合とする。次の記法を用いる;

$$(x_1, \ldots, x_N)^a = \prod_{i=1}^N (x_i)^{a_i}\,,\ a! = \prod_{i=1}^N a_i!\,,$$

$$|a| = \sum_{i=1}^N a_i, \quad \|a\| = \sum_{i=1}^N i\, a_i\,,\ \langle a \rangle = \sum_{i=1}^N (i+1) a_i = |a| + \|a\|$$

多重添字の (指数型) 部分ベル多項式 $B_{a,k}$ とは次の二重級数展開で定義される変数 $\{x_b;\ b \in S_N\}$ の多項式である;

$$\exp\left(u \sum_{b \in S_N} x_b \frac{t^b}{b!}\right) = 1 + \sum_{a \in S_N} \left\{\sum_{k=1}^{|a|} u^k B_{a,k}\right\} \frac{t^a}{a!} \tag{5}$$

ここで $t = (t_1, \ldots, t_N)$ 。$B_{a,k}$ の閉じた表現は次のようになる;

$$B_{a,k}\left(\{x_b\}\right) = \sum \frac{a!}{\prod_b m_b!} \prod_b \left(\frac{x_b}{b!}\right)^{m_b}$$

ここで和は $\sum_b m_b = k$ 且つ $\sum_b m_b b = a$ であるすべての多重添字 $\{m_b\}$ についてとる。関係 $\sum_b m_b |b| = |a|$ に注意しよう。又 $B_{a,k}\left(\{x_b\}\right)$ は $b \leq a$ (座標毎) 且つ $|b| \leq |a| - k + 1$ であるような x_b だけを含む。

性質 (3) に対応する次の関係が成り立つ;

$$B_{a,k}\left(\{a y^b x_b\}\right) = a^k y^a B_{a,k}\left(\{x_b\}\right)$$

良く知られた様にベル多項式は合成関数の微分公式 (Faà di Bruno の公式) に登場する。同様に多重添字のベル多項式は $h(x) = f(g(x))$ の形の合成関数の微分公式に登場する。$g^{(a)}(x) = (\partial/\partial x)^a g(x)$ と置こう。すると一次元の場合と同じ議論により

$$\left(\frac{\partial}{\partial x}\right)^a f(g(x)) = \sum_{k=1}^{|a|} f^{(k)}(g(x))\, B_{a,k}\left(\left\{g^{(b)}(x)\,;\, b \in S_N\right\}\right) \tag{6}$$

である事が分かる。もしここで $f(x) = \log x$ と e^x と置けば Barndorff-Nielsen and Cox (1989) で述べられた "exlog" 公式の一つの表現を得る。特に

$$\left(\frac{\partial}{\partial x}\right)^a \log g(x) = \sum_{k=1}^{|a|} \left[(-1)^{k-1}\frac{(k-1)!}{g(x)^k}\right] B_{a,k}\left(\left\{g^{(b)}(x)\right\}\right) \tag{7}$$

である。最後に次の簡単な関係が成立する;

$$Y_N(1,0,0,\ldots,0) = 1\,, \tag{8}$$

$$\left(\frac{\partial}{\partial x}\right)^a Y_N(1!x_1,\ldots,N!x_N) = \begin{cases} \frac{N!}{M!} Y_M(1!x_1,\ldots,M!x_M) & \text{if} \quad \|a\| \le N\,, \\ 0 & \text{if} \quad \|a\| > N\,, \end{cases} \tag{9}$$

ここで $M = N - \|a\|$ と置いた。

3 不一致確率

$\{p_i\}_{i\ge 1}$ を有限又は無限の確率関数とする。巾和 $\sum_{i\ge1} p_i^m$ を P_m と書く。不一致確率 r_n は次の式で与えられる

$$\sum_{i_1,i_2,\cdots,i_n} p_{i_1} p_{i_2} \cdots p_{i_n}$$

ここで和はすべての相異なる添字について取る。従って $\{r_n\}$ の母関数は次の式で与えられる

$$G(t) = 1 + \sum_{i\ge1} r_i \frac{t^i}{i!} = \prod_{i\ge1}(1 + p_i t)$$

これは Fjajolet et al. (1988) が与えた不等生起確率の誕生日問題に関する様々な母関数一つである。この母関数から r_n を巾和 $\{P_m\}$ で表す関係式が得られる:

$$\begin{aligned} G(t) &= \exp\left[\sum_{n\ge1} \log(1 + p_n t)\right] \\ &= \exp\left[\sum_{n\ge1}\left(\sum_{i=1}^{\infty}(-1)^{i-1}\frac{p_n^i}{i}\, t^i\right)\right] \end{aligned}$$

$$= \exp\left[\sum_{i=1}^{\infty}(-1)^{i-1}(i-1)!\,P_i\,\frac{t^i}{i!}\right]$$
$$= 1+\sum_{n=1}^{\infty}Y_n\left(\ldots,(-1)^{j-1}(j-1)!P_j,\ldots\right)\frac{t^n}{n!}$$

従って

$$r_n = Y_n\left(1,\ldots,(-1)^{j-1}(j-1)!P_j,\ldots,(-1)^{n-1}(n-1)!P_n\right)$$

ここでベル多項式の閉じた表現 (2) を用いれば r_n 自身の閉じた表現

$$r_n = 1+\sum_{\substack{a_1+2\cdot a_2+\cdots+n\cdot a_n=n\\ a_1\neq n}}\frac{n!}{a_1!\cdots a_n!}\prod_{i=1}^{n}\left(\frac{(-1)^{i-1}P_i}{i}\right)^{a_i} \tag{10}$$

を得る。Klotz (1979) はこの式を直接に導いている。又彼は Wisconsin の 41,208 人の誕生日のデータをもとに、この式を用いて r_n を $n=25$ 迄計算している。

ベル多項式の漸化式 (4) を用いて r_n's の漸化式を求めることが出来る:

Proposition 1 $r_0=1$ と置くと $n=1,2,\ldots$ で次の漸化式が成立する:

$$r_n = \sum_{i=1}^{n}(-1)^{i-1}\,\frac{(n-1)!}{(n-i)!}\,P_i\,r_{n-i}\,. \tag{11}$$

4 不一致確率の近似

公式 (10) は我々の問題に対する完全な解答であるが、大きな数と小さな数の積の莫大な和 (例えば r_{25} で 1958 項) となり実際の計算には不向きである。同じことは (11) 式についても言える。(もし十分な精度の多倍長演算が出来るなら (11) は不一致確率を計算するもっとも簡単な方法である、後の節を参照。) この節では不一致確率の一つの近似式を考える。$c=\max_i p_i$ と置く。一般姓を失うことなく $c<1$ とする。巾和 P_m は次の上界を持つ:

$$P_m = \sum_i p_i(p_i)^{m-1} \le c^{m-1}$$

従って

$$\begin{aligned}
&|B_{n,k}(P_1,-P_2,\ldots,(-1)^{j-1}(j-1)!P_j,\ldots)| \\
&\le \sum_{\substack{a_1+2\cdot a_2+\cdots+n\cdot a_n=n\\ a_1+a_2+\cdots+a_n=k}}\frac{n!}{a_1!\cdots a_n!}\prod_{i=1}^{n}\left(\frac{c^{i-1}}{i}\right)^{a_i} \\
&\le c^{n-k}\sum_{\substack{a_1+2\cdot a_2+\cdots+n\cdot a_n=n\\ a_1+a_2+\cdots+a_n=k}}\frac{n!}{a_1!\cdots a_n!}\prod_{i=1}^{n}\left(\frac{(i-1)!}{i!}\right)^{a_i} \\
&= c^{n-k}B_{n,k}(0!,1!,2!,\ldots) = c^{n-k}|s(n,k)|
\end{aligned} \tag{12}$$

ここで $s(n,k)$ は第1種のスターリング数である。スターリング数の定義とその部分ベル多項式による表現については Comtet (1974) を参照のこと。

r_n の部分ベル多項式による次の表現を考えよう:

$$r_n = \sum_{k=1}^{n} B_{n,k}(P_1, -P_2, \ldots, (-1)^{j-1}(j-1)!P_j, \ldots)$$

r_n を次の様に近似してみる:

$$r_{n,m} = \sum_{k=n-m+1}^{n} B_{n,k}(P_1, -P_2, \ldots, (-1)^{j-1}(j-1)!P_j, \ldots)$$

例えば

$$\begin{aligned}
r_{n,2} &= 1 - \frac{(n)_2}{2}P_2, \\
r_{n,3} &= r_{n,2} + \left[\frac{(n)_4}{8}P_2^2 + \frac{(n)_3}{3}P_3\right], \\
r_{n,4} &= r_{n,3} - \left[\frac{(n)_6}{48}P_2^3 + \frac{(n)_5}{6}P_2P_3 + \frac{(n)_4}{4}P_4\right], \\
r_{n,5} &= r_{n,4} + \left[\frac{(n)_6}{8}P_2P_4 + \frac{(n)_6}{18}P_3^2 + \frac{(n)_7}{12}P_2^2P_3 + \frac{(n)_8}{384}P_2^4\right]
\end{aligned}$$

ここで $(x)_m$ は factorial 多項式 $x(x-1)\cdots(x-m+1)$ である。

近似誤差は式 (12) を用いて次のように評価できる:

$$|r_n - r_{n,m}| \le \sum_{k=1}^{n-m} c^{n-k}|s(n,k)| = c^n \sum_{k=1}^{n-m} c^{-k}|s(n,k)| \tag{13}$$

この評価と符号無しスターリング数の母関数 (Comtet (1974, Chap. 5))、つまり

$$x(x+1)\cdots(x+n-1) = \sum_{k=1}^{n} x^k|s(n,k)|$$

を用いると次の結果が得られる:

Proposition 2

$$|r_n - r_{n,m}| \le (1+c)(1+2c)\cdots(1+(n-1)c) - \sum_{k=0}^{m-1} c^k|s(n,n-k)|. \tag{14}$$

符号無しスターリング数については次の表現と近似式が知られている (Moser and Wyman, 1958):

$$\begin{aligned}
|s(n,n)| &= 1,\\
|s(n,n-1)| &= \binom{n}{2},\\
|s(n,n-2)| &= \frac{(n)_3(3n-1)}{24},\\
|s(n,n-3)| &= \frac{(n)_2(n)_4}{48}\\
|s(n,n-4)| &= \frac{(n)_4(15n^3-30n^2+5n+2)}{5760}
\end{aligned}$$

そして $n-o(\sqrt{n}) \leq m \leq n$ に対し

$$\begin{aligned}
|s(n,m)| \simeq \binom{n}{m}\left(\frac{m}{2}\right)^{n-m} \times\\
\Big(1 + \frac{5(n-m)_2}{6m} + \frac{1}{m^2}\left\{(n-m)_3 + \frac{25(n-m)_4}{72}\right\}\\
+ \frac{1}{m^3}\left\{\frac{251(n-m)_4}{180} + \frac{5(n-m)_5}{6} + \frac{125(n-m)_6}{1296}\right\} + \cdots\Big)
\end{aligned}$$

$r_{n,m}$ による近似は n か c が小さくなければあまり良くない。例えば古典的な誕生日問題では $r_{n,4}$ に対する評価 (14) が 0.01 以下になるのは $5 \leq n \leq 23$ の範囲、誤差 $|r_n - r_{n,4}|$ 自身が 0.01 以下になるのは $5 \leq n \leq 24$ の範囲にすぎない。.

5 不一致確率の対数の近似

ベル多項式の対数の漸近展開に基ずく別の近似を与える事が出来る。この節では固定した $\{x_2, x_3, \ldots\}$ に対する

$$\log Y_N(1, 2!x_2, \ldots, N!x_N)$$

の展開を考える。この展開の $N \to +\infty$ の時の挙動に関心がある。

簡単のために次ぎの記号を用いる。集合 $\{1, 2, \ldots, K\}$ を $[K]$ と書く。記号 $\mathcal{S}(X)$ と $\mathcal{S}^2(X)$ はそれぞれ X の空でない排反部分集合族、空でない濃度 2 以上の部分集合の排反族とする。$\mathcal{I} \in \mathcal{S}(X)$ に対する合併集合 $\cup_{I\in\mathcal{I}} I$ を $\cup\mathcal{I}$ と書く。$\cup\mathcal{I} = X$ である $\mathcal{I} \in \mathcal{S}(X)$ の全体を $\mathcal{S}^*(X)$ と書く。$\mathcal{I}, \mathcal{J} \in \mathcal{S}(X)$ に対して関係 $\mathcal{I} \ll \mathcal{J}$ は、各 $I \in \mathcal{I}$ がある $J \in \mathcal{J}$ に含まれ、各 $J \in \mathcal{J}$ は高々一つの $I \in \mathcal{I}$ を含む事を意味する。もし $\cup\mathcal{I} = \cup\mathcal{J}$ $(= Y$ と置こう) であれ

ば順序関係 $\mathcal{I} \leq \mathcal{J}$ は $\mathcal{I}$ が $\mathcal{J}$ よりも細かな Y の分割である事を意味する。$\mu(\mathcal{I}, \mathcal{J})$ をこの順序関係に対する Möbius 関数とする、つまり:

$$\mu(\mathcal{I}, \mathcal{J}) = (-1)^{\#\mathcal{I}+\#\mathcal{J}} \prod_{J \in \mathcal{J}} \#\{I \in \mathcal{I};\, I \subset J\}$$

もし $\mathcal{I} \in \mathcal{S}(X)$ なら $\tau(\mathcal{I}) = \sum_{I \in \mathcal{I}} (\#I - 1)$、又 $\zeta(\mathcal{I}) = \prod_{I \in \mathcal{I}} (-1)^{\#I-1}(\#I - 1)$ と置こう。数列 $t = \{t_i\}$ に対し和 $\sum_{i \in I} t_i$ を t_I と書く。もし $f(x) = \sum_a c_a x^a$ ならば $\operatorname{mindeg} f$ は $\{|a|; c_a \neq 0\}$ の最小値を表す。更に $\xi(n) = (-1)^{n-1}(n-1)$ と置く。

Proposition 3 次ぎの形式的な展開が成り立つ:

$$\log Y_N(1, 2!x_2, \ldots, N!x_N) = \sum_{a \in S_{N-1}} C_a(N)(x_2, x_3, \ldots, x_N)^a \frac{1}{a!} \tag{15}$$

ここで C_a は次ぎの形の多項式である;

$$C_a(y) = \sum_{k=1}^{|a|} (-1)^{k-1}(k-1)! B_{a,k}\left(\left\{(y)_{\langle b \rangle};\, b \in S_{N-1}\right\}\right)$$

証明. $x = (x_1, \ldots, x_N)$ と置く。関係 (3), (7) そして (9) から

$$(\partial/\partial x)^a \log Y_N(1!x_1, 2!x_2, \ldots, N!x_N)|_{x=(1,0,\ldots,0)} = \sum_{k=1}^{|a|} (-1)^{k-1}(k-1)! B_{a,k}\left(\left\{(N)_{\|b\|} : b \in S_N\right\}\right)$$

従って次ぎの展開が成り立つ:

$$\log Y_N(1!x_1, 2!x_2, \ldots, N!x_N) = \sum_{a \in S_N} \left[\sum_{k=1}^{|a|} (-1)^{k-1}(k-1)! B_{a,k}\left(\left\{(N)_{\|b\|}\right\}\right)\right] \frac{y^a}{a!}$$

ここで $y = (x_1 - 1, x_2, \ldots, x_n)$。この展開で $x_1 = 1$ と置けば $a_1 = 0$ の項だけが残り、右辺は次ぎのようになる:

$$\sum_{a \in S_{N-1}} \left[\sum_{k=1}^{|a|} (-1)^{k-1}(k-1)! B_{(0,a),k}(\{(N)_{\|b\|}\})\right] \frac{z^a}{a!}$$

ここで $(0, a) = (0, a_1, \ldots, a_{N-1})$ そして $z = (x_2, \ldots, x_N)$。$|(0, a)| = |a|$ そして $\|(0, a)\| = \langle a \rangle$ を注意する。更に (5) より容易に

$$B_{(0,a),k}\left(\left\{x_b; b \in S_N\right\}\right) = B_{a,k}\left(\left\{x_{(0,b)}; b \in S_{N-1}\right\}\right)$$

従って証明が終る。

Proposition 4 多項式 $C_a(y)$ は次数 $\|a\|+1$ を持つ。

この主張の証明は二つの補題を用意してから行う。より正確には、和我々はまず $\deg C_a \leq \|a\|+1$ を示す。$\deg C_a = \|a\|+1$ である事は Proposition 5 で示される。

Lemma 1 各 $K \geq 1$ と $k \geq 1$ に対し関数 $F_{K,k}(t)$, $t=(t_1,\ldots,t_K)$ を

$$F_{K,k}(t) = \sum_{\mathcal{I}\in\mathcal{S}^*([K])} (-1)^{\#\mathcal{I}-1}(\#\mathcal{I}-1)\left\{\prod_{I\in\mathcal{I}}(1+t_I)-1\right\}^k \tag{16}$$

で定義する。全ての $K,k \geq 1$ で $mindeg\, F_{K,k} \geq K+k-1$ を仮定する。すると全ての $a \in S_{N-1}$ で $deg\, C_a \leq \|a\|+1$ となる。

証明. 関係 (8) と (9) から次式が成立する:

$$Y_N(1,2!x_2,3!x_2,\ldots,N!x_N) = 1+\sum_{a\in S_{N-1}}(N)_{\langle a\rangle}\frac{x^a}{a!}$$

ここで $x=(x_2,\ldots,x_N)$ である。この式と (15) から次のような C_a の母関数が得られる:

$$\log\left\{1+\sum_{a\in S_{N-1}}(y)_{\langle a\rangle}\frac{x^a}{a!}\right\} = \sum_{a\in S_{N-1}} C_a(y)\frac{x^a}{a!}$$

この母関数の左辺を直接展開し両辺を比較すると次の式が出る:

$$C_a(y) = \sum_{n=1}^{|a|}\frac{(-1)^{n-1}}{n}\sum_{b_1+\cdots+b_n=a}\left\{a!/\prod_i b_i!\right\}\left\{\prod_i (y)_{\langle b_i\rangle}\right\}$$

ここで最も内側の和は全ての (順序を考えない) $(b_1,\ldots,b_n)$ についてとる。多項式展開公式、Roman (1984, Chap. 4)、を用いると:

$$\begin{aligned} C_a(y) &= \sum_{k\geq 0}\Delta^k\{C_a(0)\}\frac{(y)_k}{k!} \\ &= \sum_{k\geq 0}\left[\sum_{n=1}^{|a|}\frac{(-1)^{n-1}}{n}\sum_{b_1+\cdots+b_n=a}\left\{a!/\prod_i b_i!\right\}\Delta^k\left\{\prod_i (0)_{\langle b_i\rangle}\right\}\right]\frac{(y)_k}{k!} \\ &\equiv \sum_{k\geq 0} C_{a,k}\frac{(y)_k}{k!}, \quad \text{仮に} \end{aligned}$$

ここで Δ は前進階差 $\Delta f(x) = f(x+1)-f(x)$ であり $\Delta^k f(0) = \Delta^k f(x)\big|_{x=0}$

各 $a \in S_{N-1}$ に多重添字 $\alpha=(\alpha_1,\ldots,\alpha_K) \in \{2,\ldots,N\}^K$, $\alpha_1 \leq \alpha_2 \leq \ldots \leq \alpha_K$, を $K=|a|$ かつ $\#\{i;\alpha_i = j+1\} = a_j$, $1 \leq j \leq N-1$ となるように対応させる事が出来

る。$\langle a\rangle = |\alpha|$ を注意する。例えば、もし $N=5$ で $a=(1,3,2,1)$ ならば、$K=7$ かつ $\alpha=(2,3,3,3,4,4,5)$ となる。この a to α の対応の下で、a の各分割 $b_1+\cdots+b_n=a$ は分割 $\mathcal{I}=\{I_1,\ldots,I_n\}\in\mathcal{S}^*([K])$ に対応し $\langle b_i\rangle=\alpha_{I_i}$ となる。更に次が分かる:

$$\sum\left\{a!/\prod_i c_i!\right\}\prod_i(y)_{\langle c_i\rangle}=n!\sum\prod_i(y)_{\alpha_{J_i}}$$

ここで左辺の和は $\{b_i\}$ の全ての置換 $\{c_i\}$ について取り、右辺は各 i で $\{\alpha_j\}_{j\in J_i}=\{\alpha_j\}_{j\in I_i}$ となる全ての分割 $\mathcal{J}=\{J_1,\ldots,J_n\}\in\mathcal{S}^*([K])$ について取る。従って、もし各 $\alpha\in S_K$ に対し

$$C^*_{\alpha,k}=\sum_{\mathcal{I}\in\mathcal{S}^*([K])}(-1)^{\#\mathcal{I}-1}(\#\mathcal{I}-1)!\,\Delta^k\left\{\prod_{I\in\mathcal{I}}(0)_{\alpha_I}\right\}\tag{17}$$

と置けば、もし $\alpha\in\{2,3,\ldots,N\}^K$ が $a\in S_{N-1}$ に対応すれば $C^*_{\alpha,k}=C_{a,k}$ となる。ここで $(x)_0\equiv 1$ と置く。

各 $\mathcal{I}\in\mathcal{S}^*([K])$ に対し $A=\prod_{I\in\mathcal{I}}(1+t_I)$ と置こう。すると次が分かる:

$$A^x=\sum_{\alpha=(\alpha_1,\ldots,\alpha_K)\in S_K}\left\{\prod_{I\in\mathcal{I}}(x)_{\alpha_I}\right\}\frac{t^\alpha}{\alpha!}$$

$\Delta^kA^x=A^x(A-1)^k$ であるから

$$(A-1)^k=\sum_{\alpha\in S_K}\Delta^k\left\{\prod_{I\in\mathcal{I}}(0)_{\alpha_I}\right\}\frac{t^\alpha}{\alpha!}$$

この関係式から $F_{K,k}$ が $\{C^*_{\alpha,k}\}$ の母関数である事が分かる、つまり:

$$F_{K,k}(t)=\sum_{\alpha\in S_K}C^*_{\alpha,k}\frac{t^\alpha}{\alpha!}$$

Proposition 4 を証明するには、各 $k>\|a\|+1$ に対し $C_{a,k}=0$ for $k>\|a\|+1$ を示せばよい。しかしながら $\|a\|=|\alpha|-K$ であるから、このことは $|\alpha|<K+k-1$ で $C^*_{\alpha,k}=0$ であればよく、更に後者は $\operatorname{mindeg}F_{K,k}\geq K+k-1$ であれば成立する。従って補題は証明された。

Lemma 2 関数系 $\{F_{K,k}\}$ は各 $K,k\geq 1$ に対し次の漸化式を満足する:

$$F_{K,k}(t)-\left\{\prod_{i=1}^K(1+t_i)-1\right\}F_{K,k-1}(t)=\tag{18}$$
$$\sum_{\mathcal{I}\in\mathcal{S}^2([K])}\zeta(\mathcal{I})P_{\mathcal{I}}(t)\left[\sum_{\mathcal{J}}\mu(\mathcal{I},\mathcal{J})F_{K-\tau(\mathcal{J}),k-1}\left(\{t_J\}_{J\in\mathcal{J}}\cup\{t_i\}_{i\notin\cup\mathcal{J}}\right)\right]$$

ここで最も内側の和は $\mathcal{I}\leq\mathcal{J}$ である様な $\mathcal{J}\in\mathcal{S}^*(\cup\mathcal{I})$ について取る。又次のように置いた:

$$P_{\mathcal{I}}(t)=\prod_{i\in\cup\mathcal{I}}t_i\times\prod_{i\in[K]\setminus\cup\mathcal{I}}(1+t_i)$$

証明. $F_{K,k}$ は対称関数であるから $F_{K,k}(\{t_i\}_{i\in I})$ の様な記法を用いてもよい。各 $L\subset[K]$ について成り立つ恒等式

$$\sum_{J\subset L,\#J\geq 1}(-1)^{\#J-1}(\#J-1)=-1$$

から、各 $I\subset[K]$ に対して:

$$\prod_{i\in I}(1+t_i)=(1+t_I)-\sum_{J\subset I,\#J\geq 2}\xi(\#J)\left\{\prod_J t_i\right\}\left\{\prod_{I\setminus J}(1+t_i)\right\}$$

従って $\mathcal{I}\in\mathcal{S}^*(K)$ に対して

$$\prod_{I\in\mathcal{I}}(1+t_I)-\prod_{i=1}^{K}(1+t_i)\ =\sum_{\mathcal{J}\in\mathcal{S}^2([K]),\mathcal{J}\ll\mathcal{I}}\zeta(\mathcal{J})P_{\mathcal{J}}(t)$$

最後の関係式を用いて (18) の左辺が次式に等しくなることが分かる:

$$\sum_{\mathcal{I}\in\mathcal{S}^2([K])}\zeta(\mathcal{I})P_{\mathcal{I}}(t)\left[\sum_{\mathcal{J}\in\mathcal{S}^*([K]),\mathcal{I}\ll\mathcal{J}}\xi(\#\mathcal{J})\left\{\prod_{J\in\mathcal{J}}(1+t_J)-1\right\}^{k-1}\right]$$

$\mathcal{I}\in\mathcal{S}^2([K])$ を一つ固定し $X=\cup\mathcal{I}$ とする。各 $\mathcal{L}\in\mathcal{S}^*(X)$ に対し

$$H(\mathcal{L})=\sum_{\mathcal{R}\in\mathcal{S}^*([K]),\mathcal{L}\ll\mathcal{R}}\xi(\#\mathcal{R})\left\{\prod_{R\in\mathcal{R}}(1+t_R)-1\right\}^{k-1}$$

と置く。更に $\bar{H}(\mathcal{U}),\mathcal{U}\in\mathcal{S}^*(X),$ を $\mathcal{U}\leq\mathcal{L}$ である $\mathcal{L}\in\mathcal{S}^*(X)$ に対する $H(\mathcal{L})$ の和とする。すると

$$\bar{H}(\mathcal{U})=F_{K-\tau(\mathcal{U}),k-1}(\{t_U\}_{U\in\mathcal{U}}\cup\{t_i\}_{i\notin\cup\mathcal{U}})$$

ここで Möbius の反転公式、Comtet (1974, Chap. 4, Suppl. 15)、を用いると:

$$H(\mathcal{L})=\sum_{\mathcal{U}\in\mathcal{S}^*(X),\mathcal{L}\leq\mathcal{U}}\mu(\mathcal{L},\mathcal{U})\bar{H}(\mathcal{U})$$

従って補題は証明された。

Proposition 4 の証明. 証明は各 $K\geq 1$ 毎に k に関する帰納法で証明する。最初に $F_{1,k}(t_1)=t_1^k$ を注意する。したgって Proposition 4 は $K=1$ と $k\geq 1$ で成立する。又 $n\geq 2$ なら $\Delta\{(0)_n f(0)\}=0$、$n=1$ ならば $=f(1)$ を注意しよう。従って (17) より $\alpha\neq 1\equiv(1,1,\ldots,1)$ であれば $C^*_{\alpha,1}=0$、そして $C^*_{1,1}=(-1)^{K-1}(K-1)!$。従って $F_{K,1}=$

$(-1)^{K-1}(K-1)!\prod_{1\le i\le K}t_i$ かつ $\mathrm{mindeg}\,F_{K,1}=K$。ここで $K\ge 2$ で $\mathrm{mindeg}\,F_{K,k-1}\ge K+k-2$ と仮定する。すると (18) の左辺の mindeg は $K+k-1$ に等しい。更に (18) の右辺の各項の mindeg は

$$\begin{aligned}\mathrm{mindeg}\left\{P_{\mathcal{I}}F_{K-\tau(\mathcal{J}),k-1}\right\} &= \mathrm{mindeg}\,P_{\mathcal{I}}+\mathrm{mindeg}\,F_{K-\tau(\mathcal{J}),k-1}\\ &\ge \#(\cup\mathcal{J})+\{K-\tau(\mathcal{J})+k-2\}=K+k-2+\#\mathcal{J}\ge K+k-1\end{aligned}$$

に等しくなる。ここで $\tau(\mathcal{J})=\#(\cup\mathcal{I})-\#\mathcal{J}$ に注意する。こうして証明が完了する。

Proposition 5 $a\in S_{N-1}$ に対する $C_{a,\|a\|+1}$ の具体的な形は次の様になる

$$C_{a,\|a\|+1}=(-1)^{|a|-1}(\langle a\rangle-1)!\,(2,3,\ldots,N)^a \tag{19}$$

この Proposition の証明のためには次の補題がいる。

Lemma 3 $\kappa(n)$, $n\ge 2$, を次式で証明する

$$\kappa(n)=\sum_{\mathcal{I}\in\mathcal{S}^*([n])\cap\mathcal{S}^2([n])}(\#\mathcal{I}-1)!\prod_{I\in\mathcal{I}}(\#I-1)$$

すると $\kappa(n)=(n-1)!$。

Lemma 3 の証明。$[n]$ の分割 $\mathcal{I}$ で $\#\{I\in\mathcal{I};\#I=i\}=x_i$, $1\le i\le n$, となるものの総数は

$$n!/\left\{\prod_i x_i!\times\prod_i(i!)^{x_i}\right\}$$

となる、Comtet (1974, Chap. 5)。従って

$$\begin{aligned}\kappa(n) &= \sum_{k=1}^{n}\sum\left[\left\{n!(k-1)!/\prod_{i=2}^{n}x_i!\right\}\prod_{i=2}^{n}\left\{\frac{i-1}{i!}\right\}^{x_i}\right]\\ &= \sum_{k=1}^{n}(k-1)!\,B_{n,k}(0,1,\ldots,n-k)\end{aligned} \tag{20}$$

ここで真中の式の最も内側の和は $\sum_{2\le i\le n}x_i=k$ かつ and $\sum_{2\le i\le n}ix_i=n$ という条件の下で取る。(1) 式から次が分かる:

$$\exp\left(u\left\{1+(t-1)e^t\right\}\right)=1+\sum_{n\ge 1}\left\{\sum_{k=1}^{n}u^kB_{n,k}(0,1,\ldots,n-k)\right\}\frac{t^n}{n!}$$

この式の両辺に e^{-u} をかけ $(0,\infty)$ 上で積分すると

$$\sum_{n\ge 2}\frac{t^n}{n}=\sum_{n\ge 1}\kappa(n)\frac{t^n}{n!}$$

となることが (20) から分かる。従って証明が完了する。

Proposition 5 の証明. これまでの議論から $F_{K,k}$ の零でない項の次数は高々 $K+k-1$ である。$G_{K,k}$ を次数が $K+k-1$ である $F_{K,k}$ の項の和とする。そこで (2) 式の両辺から次数 $K+k-1$ の項を抜き出すと次の漸化式が得られる:

$$G_{K,k}(t) = \left\{\sum_{i=1}^{K} t_i\right\} G_{K,k-1}(t) + \sum_{n=2}^{K}(-1)^{n-1}\kappa(n)\left[\sum_{I\in[K],\#I=n}\left\{\prod_{i\in I} t_i\right\} G_{K+1-n,k-1}\left(\{t_I\}\cup\{t_i\}_{i\notin I}\right)\right] \tag{21}$$

次に K と k に関する二重帰納法を用い、$K,k\geq 1$ で次の式が成立することを示そう:

$$G_{K,k}(t) = (-1)^{n-1}\phi_K(k)\left\{\prod_{i=1}^{K} t_i\right\}\left\{\sum_{i=1}^{K} t_i\right\}^{k-1} \tag{22}$$

ここで $\phi_K(k)$ は後で具体的に定められる定数である。Proposition 2 の証明から、$K\geq 1$ に対する $G_{K,1}$ は $\phi_K(1)=(K-1)!$ と置いた (22) 式の形を持つ事が分かる。もし (22) 式が $k=m-1$ と $K=1,2,\ldots,M-1$ で成立すれば、(21) 式を用いた適当な計算の後

$$G_{M,m}(t) = (-1)^{M-1}\phi_M(m)\left\{\prod_{i=1}^{M} t_i\right\}\left\{\sum_{i=1}^{M} t_i\right\}$$

である事が示せる。ここで

$$\phi_M(m) = \phi_M(m-1) + \sum_{n=2}^{M}\binom{M-1}{n-1}\kappa(n)\phi_{M+1-n}(m-1)$$

と置いた。再び Lemma 3 と最後の漸化式を用いた帰納法から $\phi_K(k)=(K+k-2)!/(k-1)!$ が分かる。従って $G_{K,k}$ が次式に等しいことが分かった:

$$(-1)^{K-1}(K+k-2)!\sum_{\alpha\in\{1,2,\ldots\}^K,|\alpha|=K+k-1}\left\{\prod_{i=1}^{K}\alpha_i\right\}\frac{t^\alpha}{\alpha!}$$

つまり $|\alpha|=K+k-1$ である $\alpha=(\alpha_1,\ldots,\alpha_K)\in\{1,2,\ldots\}^K$ に対して

$$C^*_{\alpha,k} = (-1)^{K-1}(K+k-2)!\prod_{i=1}^{K}\alpha_i$$

しかしながら対応 $\alpha\in\{2,\ldots,N\}^K\to a\in S_{K-1}$ と $C^*_{\alpha,k}=C_{a,k}$ から最後の関係式は $a\in S_{N-1}$ に対する次式の成立を意味する:

$$C_{a,\|a\|+1} = (-1)^{|a|-1}(\langle a\rangle-1)!\,(2,3,\ldots,N)^a$$

これが証明すべきことであった。

不一致確率 r_n の (15) と (19) に基ずく近似を考えよう。式 (15) で $x_i = (-1)^{i-1}P_i/i$ と置き、$C_a(n)$ をその最高次の項 $C_{a,\|a\|+1}(n)_{\|a\|+1}/(\|a\|+1)!$ で近似する。すると

$$\frac{1}{n}\log Y_n(1,-P_2,\ldots,(-1)^{i-1}(i-1)!P_i,\ldots,(-1)^{n-1}(n-1)!P_n) \tag{23}$$
$$\simeq \sum_{a\in S_{n-1}}(-1)^{\langle a\rangle-1}(\langle a\rangle-1)!\frac{(n-1)_{\|a\|}}{(\|a\|+1)!}\frac{(P_2,\ldots,P_n)^a}{a!}$$

$n\to\infty$ で $(n-1)_{\|a\|}\simeq n^{\|a\|}$ と $(P_2,\ldots,P_n)^a = \mathrm{O}(c^{\|a\|})$ であり、従って (23) 式の右辺の $\|a\|$ が大きな項は cn が小さい限り無視可能である。適当な項を選ぶことにより次ぎのような r_n の近似式を得る:

$$\begin{aligned}
\rho_{n,1} &= \exp\left\{-\frac{(n)_2}{2}P_2\right\}\\
\rho_{n,2} &= \rho_{n,1}\exp\left\{(n)_3\left[-\frac{P_2^2}{2}+\frac{P_3}{3}\right]\right\}\\
\rho_{n,3} &= \rho_{n,2}\exp\left\{(n)_4\left[-\frac{5}{6}P_2^3+P_2P_3-\frac{1}{4}P_4\right]\right\}\\
\rho_{n,4} &= \rho_{n,3}\exp\left\{(n)_5\left[-\frac{7}{4}P_2^4+3P_2^2P_3-P_2P_4+\frac{1}{5}P_5-\frac{1}{2}P_3^2\right]\right\}
\end{aligned}$$

$P_2 = 1/365$ に対する近似式 $\rho_{n,1}$ は誕生日問題で良く知られている、Feller (1968)。実際誕生日問題では $5\le n\le 100$ に対する $|r_n-\rho_{n,m}|$ の最大値は $m=1,2,3,4$ でそれぞれ 0.01034684, 0.00090218, 0.00055420 と 0.00054270 になり、極めて良い一致を示す、図 1 を参照のこと。

注意　近似 (23) はあくまで形式的なものであることを注意しよう。この近似に対する有用な誤差限界を示すのは難しい問題と思われる。又もし仮にそうした限界が存在したとしても、おそらくそれは $n\to\infty$ とともに $P_a\to 0$ であるときに始めて意味を持つ様なものであろう。もしそうだとすれば我々の関心の的であるかなり小さな n ではあまり意味が無いであろう。Arratia et al. (1989) は近似 $\rho_{n,1}$ がベルヌイ列の Chen-Stein ポアッソン近似の理論から導かれることを示し誤差 $|r_n-\rho_{n,1}|$ の一つの上界を与えた。しかしながら彼等の上界は n が大きいと一方 $\binom{n}{2}P_2$ が適当な大きさに留まる限りにおいて実際的な意味を持つ。例えば誕生日問題に対する彼等の誤差上界は $n\le 9$ に限って 0.01 以下である。

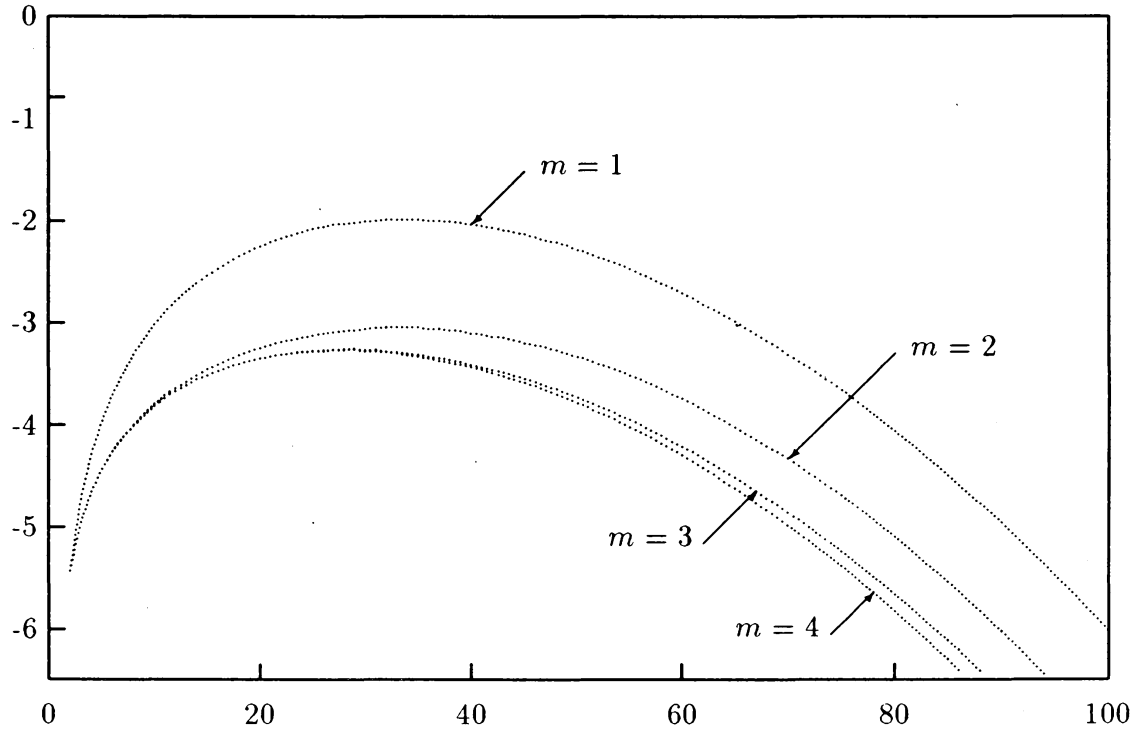

図 1 誕生日問題に対する r_n の $\rho_{n,m}$ による近似
$m = 1, 2, 3, 4$ に対する $\log_{10}|r_n - \rho_{n,m}|$ の曲線

6 同姓問題

前節の結果を同姓問題に応用してみよう。このためには日本における姓の分布を知る必要があるがこれは困難な問題である。日本には極めて多くの姓がありはたして何種類あるのか正確には分かっていない。「日本の苗字」 (1987) は確認済みの 110,867 種類の苗字を集録している。その後の調査により約 12 万種類が確認されている、丹羽 (1980)。この異常な多さの背景には歴史的、文化的、言語的な要因がある。日本の姓に関する大規模調査はこれまでに二回行われている。最初の調査は漢字による姓名の計算機処理の基礎資料を得るために日本ユニバック社によって行われた、田中 (1972)。二つの調査とも基礎資料は生命保険会社数社の顧客のレコードファイルである。その中でも第一生命保険のものがデータ数が最大であり、以下の研究の基礎資料となる、「苗字と名前」 (1987) 参照。この資料は 1986 年当時の約 8,32 万人の顧客とその 11,098,833 件の契約に基づいている (総人口 約 1 億 21,00 万人)。しかしながら公表されているデータは頻度で上位二百位迄のパーセント値 (総計で全体の約

45%) だけである、表 1 参照。

佐藤	鈴木	高橋	田中	渡辺	伊藤・東	中村	山本	小林	斎藤
1.583%	1.332%	1.132%	1.061%	1.007%	0.950%	0.864%	0.856%	0.812%	0.799%
加藤	吉田	山田	佐々木	松本	山口	木村	井上	阿・安倍	林
0.720%	0.670%	0.661%	0.590%	0.527%	0.519%	0.476%	0.470%	0.464%	0.425%
清水	森	池田	橋本	山下	石川	坂本	後藤	小川	前田
0.412%	0.384%	0.364%	0.356%	0.350%	0.333%	0.319%	0.319%	0.312%	0.310%

表 '1 30 位迄の日本の代表的名苗字
(第一生命保険会社のデータの一部)

このデータに関して幾つかの問題点があげられる。明らかにこれは無作為抽出標本では無い。又頻度は顧客数ではなく契約数で集計されており重複がある。おそらくこの二点はさほどのバイアスを生じないであろう。苗字の分布は地域的な偏りを持つことが知られているが、契約者の分布はほぼ各地域の人口に比例している。もっとも深刻な問題はこのデータが音読みで同じ姓を区別していない事であろう。一つの音読みに対し平均で約 1.7 個の漢字表記が存在するという研究がある、田中 (1972)。例えば頻度第 6 位の「いとう」には良く知られた「伊藤」「伊東」という起源の異なる二つのグループが存在する。加えて丹羽 (1981) には井東、井藤、任藤、伊統、位登、位藤、位頭、依当、依藤、威藤、夷藤、居藤、揖藤、為藤、藺藤等の例があげられている (これらはおそらく伊藤・伊東の変形であろう)。しかしながら他にどうしようもないという理由から、音読みで同じ姓を同一視する。

表 2 にユニバックの調査から得られた累積頻度を示す、田中 (1972)。このデータは第百生命の顧客 715,815 人のデータファイルから得られた。これから分かるように苗字の分布は仮に極めて稀なものを除いたとしても極めて長い裾を持つ。

順位	50	100	200	300	500	1000	2000	3000	5000	10000	20000	25000
A	27.81	37.21	48.40	55.35	63.80	74.59	83.61	87.89	92.30	96.60	99.21	99.91
B	25.27	34.46	45.28	52.31	61.75	74.95	87.01	92.58	97.26	99.70	99.98	99.99

表 2 日本の姓の累積頻度
第百生命データ A(%) と推定値 B(%)

従って通常の分布をこのデータに当てはめることは困難である。幾つかの試行錯誤の結果次の形の関数が少なくとも順位 200 位迄の範囲で比較的良くあてはまりる事が分かった:

$$f(n) = \frac{d}{n^a}$$

ここで a は約 0.7。もし $a > 1$ ならば正規化すればこれはゼータ分布 (Zipf 分布) になる。ここで Zipf 分布が物のサイズ・占有率の近似理論分布になると言う Hill (1974) の研究を思い出そう。しかしながら、$a \leq 1$ ならばこれは発散級数になるため、収束因子 c^n, $c < 1$ をかけて収束級数を得る。長い裾を持つためには c は 1 に近くなければならない。結果として我々は次のような関数を当てはめることにした。

$$f(n : a, b, c, d) = d\frac{c^n}{(n+b)^a}, \quad n = 1, 2, 3, \cdots. \tag{24}$$

和 $\sum_{n=0}^{\infty} f(n : a, b, c, 1)$ で定義される関数 $\Phi(c, a, b)$ は一般化されたリーマンのゼータ関数と呼ばれる、Gradshteyn and Ryzhik (1980)。そして無限和 $\Phi(c, a, b)$ を計算することが困難なため、日本の苗字の概数は 12 万種という情報を基に和の範囲を限って次の密度関数を考える:

$$p(n : a, b, c) = \xi(a, b, c)^{-1}\frac{c^n}{(n+b)^a}, \quad 1 \leq n \leq 120,000 \tag{25}$$

ここで $\xi(a, b, c)$ は正規化定数

$$\xi(a, b, c) = \sum_{n=1}^{120,000} \frac{c^n}{(n+b)^a}$$

打ちきりによる残差は実際上無視可能である。次に非線型最少自乗法による当てはめを行った。三母数密度関数 (25) を用いた当てはめは、しかしながら和 ξ の計算に時間がかかりすぎることと、得られた値の不正確さとのために失敗した。四母数関数 (24) を当てはめると $a = 0.9474$, $b = 5.798$, $c = 1.002$ そして $d = 0.09464$ という値が得られた。この値は決定係数 $R^2 = 99.38\%$ という良い当てはめを示すが、$c > 1$ であるため 200 位以上に補外すると意味が無くなる。結果として我々は適当に選んだ $c < 1$ に対し関数 (24) を三母数 (a, b, d) で当てはめ、得られたパラメータ値から計算した (24) の和が出来るだけ 1 に近くなるように逆に c を定めることにした。このようにして我々は最終的な推定値 $a = 0.7570$, $b = 3.648$, $c = 0.9996$ を得た。決定係数は $R^2 = 99.27\%$。次にこの推定値を関数 (25) に代入し $\xi(a, b, c) = 19.80427$ を得、更に補外により日本の姓の頻度を推定した。表 2 に推定累積頻度を示した。以上の計算及び以下の計算に際しては非線型最少自乗法を除くすべての数値

計算を UBASIC を用いて行った。UBASIC は金沢大学数学教室の木田祐司氏により開発された最大 2,600 桁 (第 8 版) の固定小数点実数が扱える MSDOS 用のベーシック言語である。

以上の準備を基に同姓確率 R_n を求めてみよう。先ず漸化式 (11) を直接用いる。結果は表 3 にまとめた、図 2 も参照されたい。一致確率が 50% を越えるのは $n = 27$ の時である。その時 $R_{27} = 51.153\%$。同様に典型的な場合は $R_{18} = 27.327\%$, $R_{38} = 75.332\%$, $R_{41} = 80.250\%$, $R_{50} = 90.727\%$, $R_{57} = 95.289\%$, そして $R_{71} = 99.028\%$。図 3 に不一致確率 r_n を $\rho_{n,m}$, $m = 1, 2, 3, 4$, で近似した時の近似誤差を図示した。近似の程度はやはり非常に良い。$\rho_{n,3}$ に見られる尖りは r_3 と $\rho_{n,3}$ がたまたまその当りで非常に近くなったために生じたものである。

R_5	R_{10}	R_{15}	R_{20}	R_{25}	R_{30}	R_{35}	R_{40}	R_{45}	R_{50}
2.14	9.14	19.81	32.60	45.95	58.60	69.65	78.70	85.66	90.73
R_{55}	R_{60}	R_{65}	R_{70}	R_{75}	R_{80}	R_{85}	R_{90}	R_{95}	R_{100}
94:24	96.56	98.02	98.90	99.41	99.70	99.85	99.93	99.97	99.99

表 3. 同姓確率 (%)

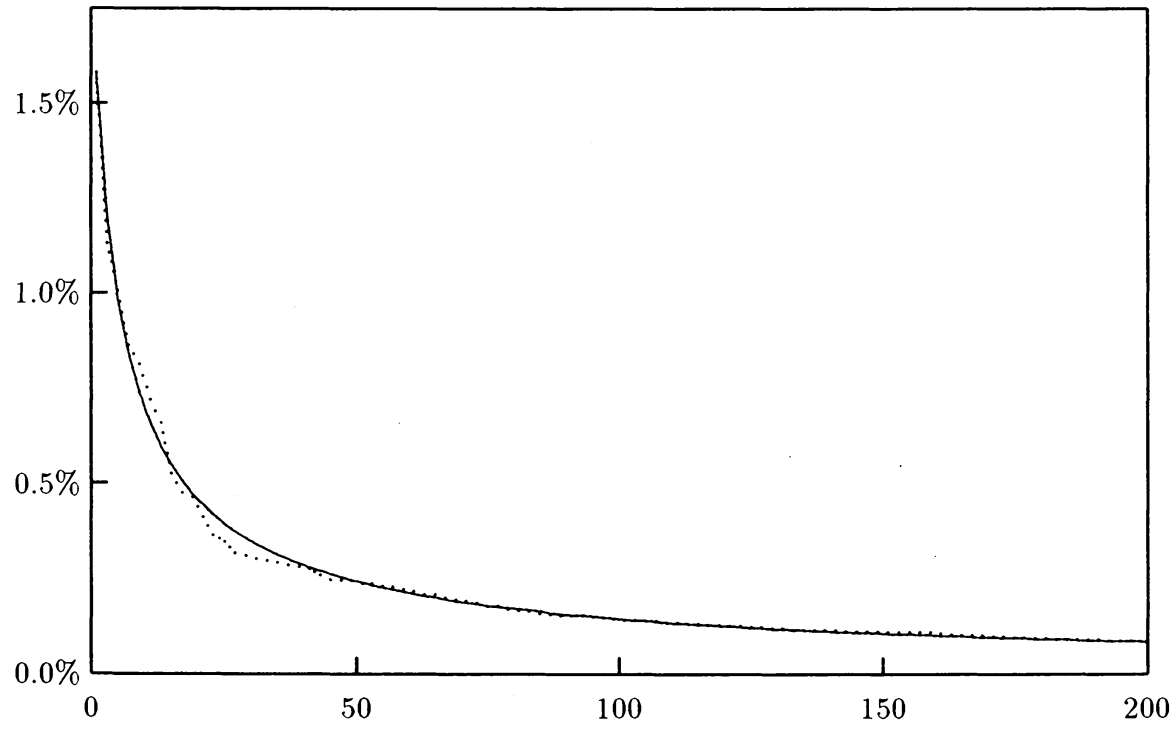

図 2　200 位迄の姓の頻度と当てはめられた曲線

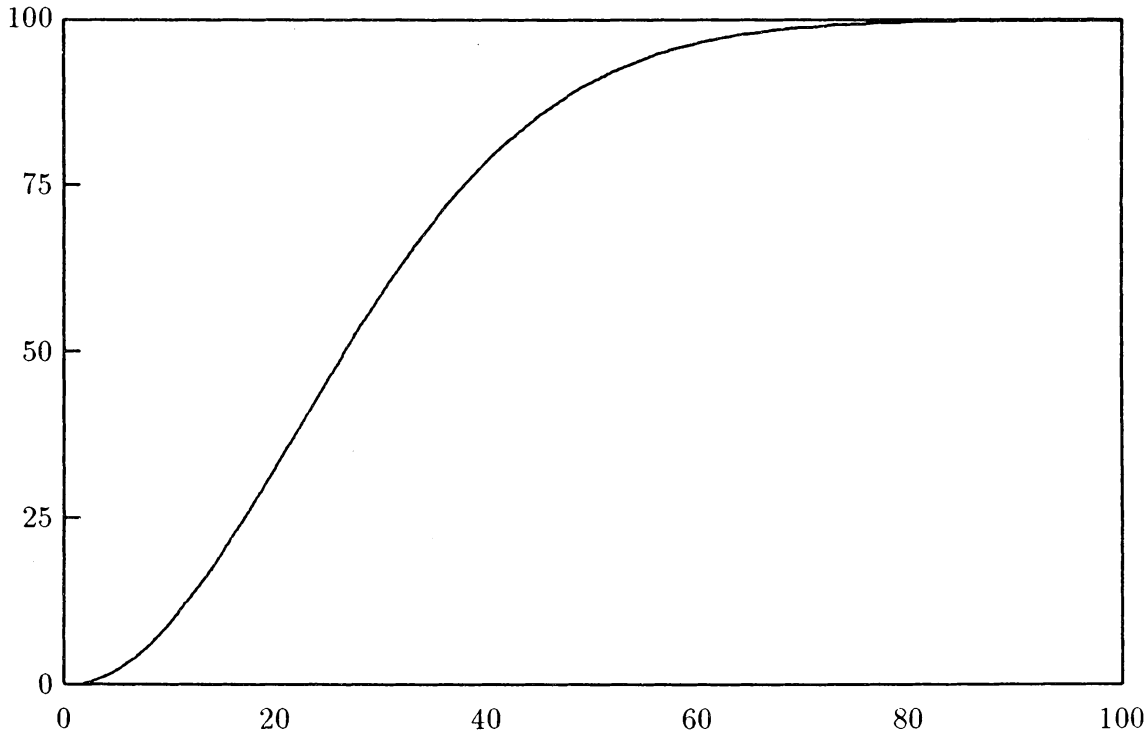

図 3.　$N = 2$ から 100 迄の誕生日確率

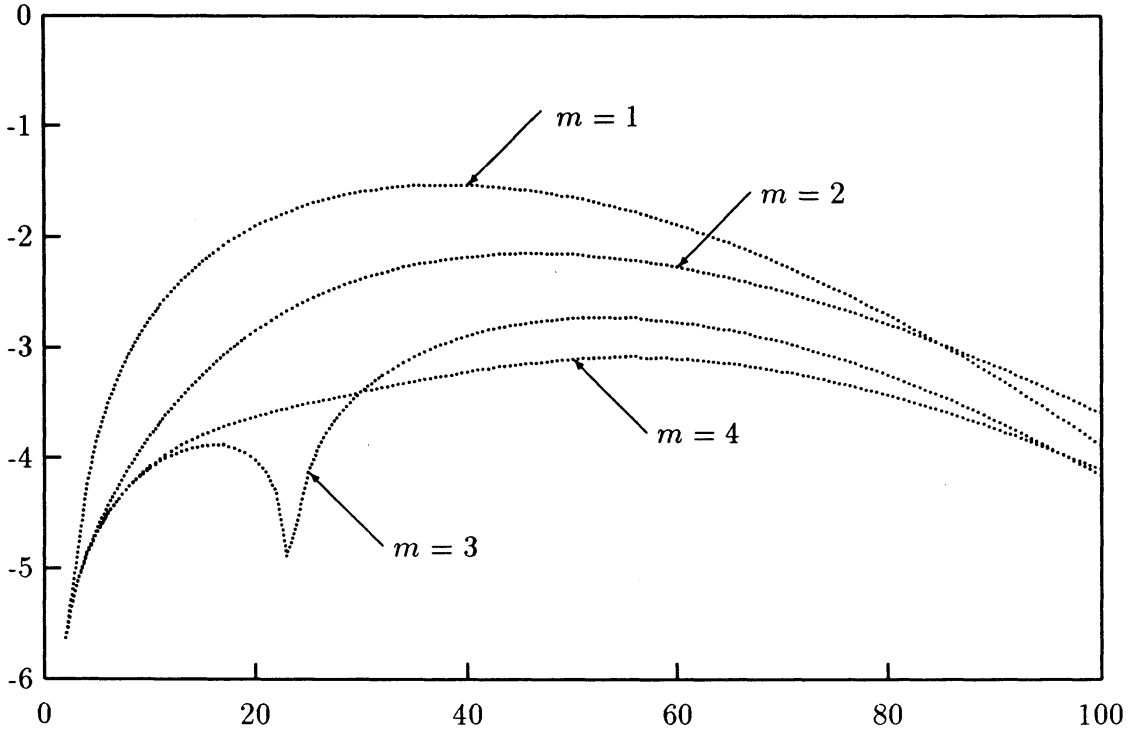

図 3. 誕生日問題: r_n の $\rho_{n,m}$ による近似
$m = 1, 2, 3, 4$ に対する $\log_{10}|r_n - \rho_{n,m}|$ の曲線

7 “真の”誕生日問題

二つ目の応用として日本における真の誕生日一致確率を推定してみよう。実際の誕生日の分布は一様とは程遠く、又世代毎に大きな変化を見せる。例えば、1900 年の日本の最大月間出生数は 154,114 人 (一月)、最少は 85,469 人 (六月) であり、その比は 180:100。一方 1980 年では最大は 134,734 人 (八月)、最少は 116,152 人 (二月) で、その比は 116:100。更に年間出生数も世代毎に大きく変化する。特定の誕生日に対する好み・忌避による虚偽の誕生日の申告も見逃せない。Klotz (1979) は医者の都合が誕生日の分布を偏らせる重要な要因だと指摘している。

勿論日本人の誕生日の日別分布の資料があるとは思えず、既存の部分的資料から推定せざるを得ない。以下の作業の基礎資料は 1900(5)1940,1947,1950(5)1980, そして 1982(1)1987 の月別出生数 (Vital Statistics 1987, JAPAN (1988)) と 1900(1)1988 の各年度に生まれた人口の 1988 年における生存数 (Japan Statistical Yearbook 1989、(1990)) である。推定は次ぎ

のような単純な区分的線型補間で行った。$M(i), 1 \leq i \leq 12$ をある年の月別出生数とする。$M(i)$ を年間総出生数と月の日数で割り比 $m(i)$ を得る。$d(i)$ を月 i の真中の時点とする。点 $P(i) = (d(i), m(i)), 0 \leq i \leq 13$, を区分的に線分で結ぶ、ここで $P(0) = (d(12) - 365, m(12))$ と $P(13) = (d(1) + 365, m(1))$ と置いた。正規化を施してから日別の誕生日比率 $n(j), 1 \leq j \leq 365$ を求める。閏年の 2 月 29 日は無視した。月別出生数が得られない年については、月別出生数が得られた前後の年の推定誕生日比率を線形補間して得た。最後に 1988 年時点における各世代の生存数をかけて 365 日毎の誕生数を求めた。このようにして得られた日別出生数はやはりかなりの変動を見せた。最大は 0.343% (1 月 16 日)、最少は 0.236% (6 月 15 日)。1/365=0.274% を基準にした比率は 125:100:86。

以上の作業をもとに真の誕生日確率を求めてみよう。しかしながら誕生日比率が一様とは程遠いにも関わらず、誕生日確率は等確率の仮定から求めたそれとごく僅かな違いしか見せなかった。最大の差は $n = 27$ における 0.370%。Klotz (1979) も同じ現象を報告している。一致確率は等確率の場合に最小になることが知られている、Bloom (1973)、Munford (1977)。Munford はこの事実が、何故実際には誕生日の一致がより頻繁に観察されるのかを説明すると述べている。しかし我々及び Klotz の例はこのことと矛盾する。おそらく Munford は個人的な経験を誤って一般化したのであろう。

一致確率の安定性は次ぎのように説明することが出来る。これまでの議論から不一致確率 r_n は既に $\rho_{n,1}$ でかなりの程度に近似可能な事が分かっている。そして $\rho_{n,1}$ は P_2 だけで決る。$p_i = (1 + \epsilon_i)/365$ と置こう。$\sum_i \epsilon_i = 0$ であり

$$\sum_i {p_i}^2 - \sum_i \left(\frac{1}{365}\right)^2 = \frac{1}{365^2} \sum_i {\epsilon_i}^2$$

つまり誕生日の比率の 1/365 からのずれは P_2 の値の 1/365 から $(1 + \sigma^2)/365$ へのずれに相当する、ここで σ^2 は $\{\epsilon_i\}$ の分散である。従って $\rho_{n,1}$ の値の相対的なずれは近似的に次ぎの様になる:

$$\frac{(n)_2}{2} \frac{\sigma^2}{365}$$

我々の例では $\sigma^2/365$ は 0.00003 であり、従って誕生日確率はほとんど変わらないことになる。日別誕生数の推定に際し我々の用いた方法以外の方法も考えられるであろうが、こうした事情は誕生日確率そのものはほとんど影響を受けない事を示している。

謝辞

この研究に当っては慶応大学の渋谷、前島両教授から関連文献に関する御教授を頂いた。ここに記して感謝したい。

文献

Arratia, R., L. Goldstein, and L. Gordon (1989) Two moments suffice for Poisson approximations: the Chen-Stein method. *Annal. Prob.*, **17**, 9–25.

Barndorff-Nielsen, O.E. and D.R. Cox (1989) *Asymptotic Techniques for Use in Statistics*, Chapman and Hall London, New York.

Bloom, D. M. (1973) A birthday problem, *American Math. Monthly*, **80**, 1141–1142.

Bolotonikov, Yu. V. (1968) Limiting processes in a model of distribution of particles into cells with unequal probabilities, *Theory Prob. Appl.*, **13**, 504–511.

Chistyakov, V. P. and Viktorova, I. I. (1965) Asymptotic normality in a problem of balls when probabilities of falling into different boxes are different, *Theory Prob. Appl.*, **10**, 149–154.

Comtet, L. (1974) *Advanced Combinatorics*, D. Reidel Publishing co., Dordrecht.

第一生命広報部(編) (1987) 苗字と名前. 恒友出版

Fang, K.-T. (1987) Occupancy problems, in *Encyclopedia of Statistical Sciences*, (eds. S. Kotz and N. L. Johnson), Vol. 6, 402–406, Wiley, New York.

Feller, W. (1968) *An Introduction to Probability Theory and its Applications*, Vol. 1, Wiley, New York.

Flajolet, P., D. Gardy, and L. Thimonier (1988) Probabilistic Languages and Random Allocations, in *Lecture Notes in Computer Sciences*, Vol. 317, 239–253, Springer Verlag. See also the paper of the same authors "Birthday paradox, coupon collectors, caching algorithms and self-organizing search" which will appear in *Discrete Applied Mathematics* (1991).

Gradshteyn, I. S. and Ryzhik, I. M. (1980) *Tables of Integrals, Series, and Products.* Academic Press, Orland.

Hill, B. M. (1974) The rank-frequency form of Zipf's law, *J. Amer. Statist. Assoc.*, **69**,

1017–1026.

Klotz, J. (1979) *The birthday problem with unequal probabilities*, Technical Report No. 59, Department of Statistics, University of Wisconsin.

Kolchin V. F., Sevast'yanov, B. A. and Chistyakov, V. H. (1987) *Random Allocation*, (translation ed. A. V. Barakrishna), Wistons and sons, Washington D. C.

Kotz. S. and Johnson, N. L. (1977) *Urn Models and Their Applications*, Wiley, New York.

Management and Coordination Agency (ed.) (1990) *Japan Statistical Yearbook 1989*, Statistics Bureau, Management and Coordination Agency.

Ministry of Health and Welfare (ed.) (1988) *Vital Statistics 1987, JAPAN*, Volume 1, Statistics and Information Department, Minister's Secretariat, Ministry of Health and Welfare.

Moser, L. and Wyman, M. (1958) Asymptotic development of the Stirling numbers of the first kind. *J. London Math. Soc.*, **33**, 133–146.

Munford, A. G. (1977) A note on the uniformity assumption in the birthday problem, *American Statistician*, **31**, 119.

Nishimura, K. and Sibuya M. (1988) Occupancy with two types of balls, *Annals Inst. Statist. Math.*, **40**, 77-91.

丹羽基二 (編) (1978) 日本の苗字、表記編・表音編. 日本経済新聞社刊.

丹羽基二 (1980) 姓氏の語源. 角川書店

Roman, S. (1984) *The Umbral Calculus.* Academic Press, Orland.

田中康仁 (1972) 日本人の姓と名の統計. 言語生活, No.**254**, 72–79, 筑摩書房.