Ann. Inst. Statist. Math.
Vol. 44, No. 3, 479–499 (1992)

APPROXIMATIONS TO THE BIRTHDAY PROBLEM WITH UNEQUAL OCCURRENCE PROBABILITIES AND THEIR APPLICATION TO THE SURNAME PROBLEM IN JAPAN*

SHIGERU MASE

*Faculty of Integrated Arts and Sciences, Hiroshima University,
Naka-ku, Hiroshima 730, Japan*

(Received July 4, 1990; revised September 2, 1991)

Abstract. Let $X_1, X_2, \dots, X_n$ be iid random variables with a discrete distribution $\{p_i\}_{i=1}^m$. We will discuss the coincidence probability R_n, i.e., the probability that there are members of $\{X_i\}$ having the same value. If $m = 365$ and $p_i \equiv 1/365$, this is the famous birthday problem. Also we will give two kinds of approximation to this probability. Finally we will give two applications. The first is the estimation of the coincidence probability of surnames in Japan. For this purpose, we will fit a generalized zeta distribution to a frequency data of surnames in Japan. The second is the *true* birthday problem, that is, we will evaluate the birthday probability in Japan using the actual (non-uniform) distribution of birthdays in Japan.

Key words and phrases: Birthday problem, coincidence probability, non-uniformness, Bell polynomial, approximation, surname.

1. Introduction

It is frequently observed that, even within a small group, there are people with the same surname. It may not be considered so curious to find persons with the same surname in a group compared to those with the same birthday. But if we consider the variety of surnames and the relatively small portions of each surname in some countries, this fact becomes less trivial than is first seen. The unusualness is most clearly seen in author's country, Japan, where there exist at least 120,000 surnames, and the cumulative percentage of surnames till 5,000th rank is still 92.3%; that is, there are extremely many surnames if rare names are not taken account of. Therefore it seems an interesting problem to know the probability that there exist persons having the same surname in a given group of n persons.

This problem, which we have already noticed and would like to call *the surname problem*, is nothing but the birthday problem with unequal probabilities of

* This research is supported in part by Grant-in-Aid for Scientific Research of the Ministry of Education, Science and Culture under the contact number 01540141 and 02640057.

occurrence. The birthday problem is well-known since its implication contradicts our intuition considerably. Also it is generalized in many ways, see, Feller (1968), Johnson and Kotz (1977), Kolchin *et al.* (1978) and Fang (1985). One recent example is Nishimura and Sibuya (1988) where authors considered the coincidence probability of birthdays between two groups.

But in almost all cases, the uniform occurrence probabilities are assumed and papers which consider unequal occurrence probabilities seem to be scarce. Early examples are Chistyakov and Viktorova (1965) and Bolotonikov (1968). Their interests were in the limiting forms of occupancy distributions. Recently Flajolet *et al.* (1988) showed that generating functions for several combinatorial problems can be derived in a unified and elegant manner using the concept of the shuffling operation of formal languages, see also Flajolet *et al.* (1991). Among others, they derived generating relations for various basic probabilities of the birthday problem with unequal occurrence probabilities.

As to practical evaluations of coincidence probabilities for non-uniform occurrence probabilities, we know only one paper of Klotz (1979). In his interesting paper, Klotz remarked that the true distribution of birthdays is never uniform and derived the formula of non-coincidence probability r_n for non-uniform birthday rates. Also he calculated r_n up to $n = 25$ using his formula directly for data of birthdays for 41,208 Wisconsin residents who died in 1975.

Now we define the surname problem formally as follows: Let $X_1, X_2, \dots, X_n$ be independent random variables which have an identical (infinite or finite) discrete distribution $\boldsymbol{P}(X = j) = p_j$ for $j = 1, 2, \dots$. What is the coincidence probability R_n that at least two of X_i have the same value? In this paper, we first express non-coincidence probabilities $r_n = 1 - R_n$ using Bell polynomials and power sums $P_a = \sum_{i\geq 1} p_i^a$, which is equivalent to the formula of Klotz. From this expression we derive a recurrence relation for r_n. Also we give two types of approximations of non-coincidence probabilities. The second approximation is based on a formula of asymptotic expansion of logarithm of Bell polynomials and shows very fine agreements.

Finally, we try to estimate the surname coincidence probabilities in Japan. This is a difficult problem since there are extremely many surnames in Japan and there is no complete global study of the distribution of Japanese surnames. We use the data of insurants of Daiiti Life Insurance Co. which consists of percentages of surnames up to 200th rank. We fit a generalized zeta distribution to this data by a non-linear regression and estimate percentages of surnames over 200th rank by extrapolation. From the estimated distribution we calculate the coincidence probabilities of surnames in Japan. As expected, these probabilities are very large even for rather small n. For example, the probability exceeds 50% already at $n = 27$ and exceeds 90% even at $n = 50$.

Also we apply our formulas to the *true* birthday problem in Japan. For this purpose, we try to estimate the actual distribution of birthdays of Japanese (who lived in 1988 and were born from 1900 to 1987) from monthly birth number data. The estimated distribution shows in fact a considerable fluctuation. However, it is shown that coincidence probabilities differ very slightly from those calculated based on the uniform birth rate assumption.

2. Bell polynomials

In the following, we will use Bell polynomials extensively. So we summarize the necessary facts first. For details, see Comtet (1974) or Roman (1984). (Exponential) partial Bell polynomials $B_{n,k}(x_1,\ldots,x_{n-k+1})$ in variables $x_1,x_2,\ldots$ are defined by the formal double series expansion;

$$\exp\left(u\sum_{m\geq 1} x_m \frac{t^m}{m!}\right) = 1+\sum_{n\geq 1}\left\{\sum_{k=1}^{n} u^k B_{n,k}(x_1,x_2,\ldots)\right\}\frac{t^n}{n!}. \tag{2.1}$$

(Exponential) complete Bell polynomials $Y_n(x_1,\ldots,x_n)$ are defined by

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

that is, $Y_n = \sum_{k=1}^{n} B_{n,k}$ and $Y_0 = 1$. The precise form of $B_{n,k}$ is

$$B_{n,k} = \sum \frac{n!}{a_1!\cdots a_n!}\prod_{i=1}^{n}\left(\frac{x_i}{i!}\right)^{a_i}, \tag{2.2}$$

where the summation is taken over all $(a_1,a_2,\ldots,a_n)$ such that $\sum_{i=1}^{n} a_i = k$ and $\sum_{i=1}^{n} ia_i = n$. In particular, $B_{n,n} = x_1^n$. Following homogeneous properties are immediate:

$$\begin{aligned} B_{n,k}(abx_1,\ldots,ab^jx_j,\ldots) &= a^kb^nB_{n,k}(x_1,\ldots,x_j,\ldots),\\ Y_n(bx_1,b^2x_2,\ldots,b^nx_n) &= b^nY_n(x_1,\ldots,x_n). \end{aligned} \tag{2.3}$$

It is known that Bell polynomials have the recurrence relation

$$Y_n(x_1,\ldots,x_n) = \sum_{i=1}^{n}\binom{n-1}{i-1}x_iY_{n-i}(x_1,\ldots,x_{n-i}), \tag{2.4}$$

see Roman ((1984), Chapter 4.1.8).

Also we will need the concept of multi-indexed Bell polynomials. Although the author cannot find the definition of this concept in the literature, it is a direct generalization of the ordinary one. Let S_N be the set of multi-indices $\boldsymbol{a} = (a_1,\ldots,a_N) \neq (0,\ldots,0)$. We use following notations;

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

$$|\boldsymbol{a}| = \sum_{i=1}^{N} a_i, \qquad \|\boldsymbol{a}\| = \sum_{i=1}^{N} ia_i, \qquad \langle\boldsymbol{a}\rangle = \sum_{i=1}^{N}(i+1)a_i = |\boldsymbol{a}| + \|\boldsymbol{a}\|.$$

The (partial exponential) multi-indexed Bell polynomials $B_{\boldsymbol{a},k}$ are polynomials in variables $\{x_{\boldsymbol{b}}; \boldsymbol{b} \in S_N\}$ defined by the following formal double series expansion in $\boldsymbol{t} = (t_1, \dots, t_N)$:

$$\exp\left(u \sum_{\boldsymbol{b}\in S_N} x_{\boldsymbol{b}} \frac{\boldsymbol{t}^{\boldsymbol{b}}}{\boldsymbol{b}!}\right) = 1 + \sum_{\boldsymbol{a}\in S_N} \left\{\sum_{k=1}^{|\boldsymbol{a}|} u^k B_{\boldsymbol{a},k}\right\} \frac{\boldsymbol{t}^{\boldsymbol{a}}}{\boldsymbol{a}!}. \tag{2.5}$$

We can show that the closed expression of $B_{\boldsymbol{a},k}$ is

$$B_{\boldsymbol{a},k}\left(\{x_{\boldsymbol{b}}\}\right) = \sum \frac{\boldsymbol{a}!}{\prod_{\boldsymbol{b}} m_{\boldsymbol{b}}!} \prod_{\boldsymbol{b}} \left(\frac{x_{\boldsymbol{b}}}{\boldsymbol{b}!}\right)^{m_{\boldsymbol{b}}},$$

where the summation is taken over all multi-indices $\{m_{\boldsymbol{b}}\}$, $|\boldsymbol{b}| \leq |\boldsymbol{a}|$, such that $\sum_{\boldsymbol{b}} m_{\boldsymbol{b}} = k$ and $\sum_{\boldsymbol{b}} m_{\boldsymbol{b}} \boldsymbol{b} = \boldsymbol{a}$. Note that $\sum_{\boldsymbol{b}} m_{\boldsymbol{b}} |\boldsymbol{b}| = |\boldsymbol{a}|$. Also $B_{\boldsymbol{a},k}\left(\{x_{\boldsymbol{b}}\}\right)$ depends only on those $x_{\boldsymbol{b}}$ with $\boldsymbol{b} \leq \boldsymbol{a}$ (coordinate-wise) and $|\boldsymbol{b}| \leq |\boldsymbol{a}| - k + 1$.

Corresponding to (2.3), we have the relation $B_{\boldsymbol{a},k}(\{a\boldsymbol{y}^{\boldsymbol{b}} x_{\boldsymbol{b}}\}) = a^k \boldsymbol{y}^{\boldsymbol{a}} \times B_{\boldsymbol{a},k}(\{x_{\boldsymbol{b}}\})$. As is well-known, Bell polynomials play an important role in differentiation of composite functions. Analogously multi-indexed Bell polynomials are useful in differentiation of composite functions of the form $h(\boldsymbol{x}) = f(g(\boldsymbol{x}))$. Let $g^{(\boldsymbol{a})}(\boldsymbol{x}) = (\partial/\partial\boldsymbol{x})^{\boldsymbol{a}} g(\boldsymbol{x})$. Then we can show that

$$\left(\frac{\partial}{\partial \boldsymbol{x}}\right)^{\boldsymbol{a}} f(g(\boldsymbol{x})) = \sum_{k=1}^{|\boldsymbol{a}|} f^{(k)}(g(\boldsymbol{x})) B_{\boldsymbol{a},k}\left(\left\{g^{(\boldsymbol{b})}(\boldsymbol{x}); \boldsymbol{b} \in S_N\right\}\right). \tag{2.6}$$

The proof is almost the same as that of the case $N = 1$, i.e., Faà di Bruno formula's, see Comtet ((1974), Chapter 3). If we let $f(x) = \log x$ or e^x in this formula we get one expression of the exlog relation discussed in Barndorff-Nielsen and Cox (1989). In particular,

$$\left(\frac{\partial}{\partial \boldsymbol{x}}\right)^{\boldsymbol{a}} \log g(\boldsymbol{x}) = \sum_{k=1}^{|\boldsymbol{a}|} \left[(-1)^{k-1} \frac{(k-1)!}{g(\boldsymbol{x})^k}\right] B_{\boldsymbol{a},k}\left(\left\{g^{(\boldsymbol{b})}(\boldsymbol{x})\right\}\right). \tag{2.7}$$

Finally we need following easy properties:

$$Y_N(1, 0, 0, \dots, 0) = 1, \tag{2.8}$$

$$\left(\frac{\partial}{\partial \boldsymbol{x}}\right)^{\boldsymbol{a}} Y_N(1!x_1, \dots, N!x_N) = \begin{cases} \dfrac{N!}{M!} Y_M(1!x_1, \dots, M!x_M) & \text{if } \|\boldsymbol{a}\| \leq N, \\ 0 & \text{if } \|\boldsymbol{a}\| > N, \end{cases} \tag{2.9}$$

where $M = N - \|\boldsymbol{a}\|$.

3. Non-coincidence probabilities

Let $\{p_i\}_{i\geq 1}$ be a given (finite or infinite) probability distribution. Power sums $\sum_{i\geq 1} p_i^m$ are denoted by P_m. It is easy to see that the probability r_n of non-coincidence is equal to $\sum_{i_1,i_2,\ldots,i_n} p_{i_1}p_{i_2}\cdots p_{i_n}$, where the summation is taken over unordered mutually distinct indices. Therefore, the exponential generating function of $\{r_n\}$ is

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

This is the simplest case of a family of generating functions due to Flajolet *et al.* (1988) for various probabilities of the birthday problem with unequal occurrence probabilities, see also Flajolet *et al.* (1991). From this generating function we can get an expression of r_n's in terms of power sums $\{P_m\}$ as follows.

$$\begin{aligned} G(t) &= \exp\left[\sum_{n\geq 1}\log(1+p_n t)\right] \\ &= \exp\left[\sum_{n\geq 1}\left(\sum_{i=1}^{\infty}(-1)^{i-1}\frac{p_n^i}{i}t^i\right)\right] \\ &= \exp\left[\sum_{i=1}^{\infty}(-1)^{i-1}(i-1)!P_i\frac{t^i}{i!}\right] \\ &= 1+\sum_{n=1}^{\infty} Y_n\left(\ldots,(-1)^{j-1}(j-1)!P_j,\ldots\right)\frac{t^n}{n!}. \end{aligned}$$

Therefore

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

If we use the closed expression (2.2) of Bell polynomials, we can get the following explicit expression of r_n from the last relation:

$$r_n = 1 + \sum_{\substack{a_1+2\cdot a_2+\cdots+n\cdot a_n=n \\ a_1\neq n}} \frac{n!}{a_1!\cdots a_n!}\prod_{i=1}^{n}\left(\frac{(-1)^{i-1}P_i}{i}\right)^{a_i}. \tag{3.1}$$

This is the expression given by Klotz (1979). He derived it directly without using Bell polynomials. Also he used this expression to calculate r_n up to $n=25$ using data of birthdays of 41,208 Wisconsin residents.

Using the recurrence relation (2.4) we can get the following recurrence relation for r_n's immediately.

PROPOSITION 3.1. *If we set $r_0 = 1$, then for $n = 1, 2, \ldots$*

$$r_n = \sum_{i=1}^{n}(-1)^{i-1}\frac{(n-1)!}{(n-i)!}P_i\, r_{n-i}. \tag{3.2}$$

4. Approximation of non-coincidence probabilities

Although the formula (3.1) is a complete answer to our problem, it is in general, a sum of numerous terms (for example, 1,958 summands for r_{25}) each of which are products of big combinatorial numbers and small numbers $\{P_m\}$. Also the relation (3.2) involves products of big combinatorial numbers and small numbers and is not suitable to actual numerical evaluations. (If we can make calculations with sufficiently high order precision, (3.2) is the easiest method to calculate non-coincidence probabilities, see later sections.)

In this section, we consider one approximation to non-coincidence probabilities. Let $c = \max_i p_i$. We can assume $c < 1$ without loss of generality. The m-th power sum P_m can be bounded as $P_m = \sum_i p_i (p_i)^{m-1} \leq c^{m-1}$. Therefore

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

Here $s(n,k)$ is the Stirling number of the first kind. As to the definition of Stirling numbers and their expression in terms of partial Bell polynomials, see Comtet's book (1974).

Note that the expression of r_n by partial Bell polynomials

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

Now we try to approximate r_n by

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

For example,

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

where $(x)_m$ stands for the factorial polynomial $x(x-1)\cdots(x-m+1)$.

We can bound approximation errors using (4.1) as follows

$$|r_n - r_{n,m}| \leq \sum_{k=1}^{n-m} c^{n-k}|s(n,k)| = c^n \sum_{k=1}^{n-m} c^{-k}|s(n,k)|. \tag{4.2}$$

If we use this estimate and the generating relation of signless Stirling numbers of the first kind, that is,

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

see Comtet ((1974), Chapter 5), we can get the following result.

PROPOSITION 4.1.

$$|r_n - r_{n,m}| \leq (1+c)(1+2c)\cdots(1+(n-1)c) - \sum_{k=0}^{m-1} c^k|s(n,n-k)|. \tag{4.3}$$

As to the signless Stirling numbers, following expressions and approximations are known, see Moser and Wyman (1958);

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

and, for $n - o(\sqrt{n}) \leq m \leq n$,

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

Approximations by $r_{n,m}$ are poor unless either n or c is small. For the birthday problem, the error bound in (4.3) for $r_{n,4}$ is less than 0.01 only for $5 \leq n \leq 23$ and the error $|r_n - r_{n,4}|$ itself is less than 0.01 only for $5 \leq n \leq 24$.

5. Approximation of logarithms of non-coincidence probabilities

We can give another type of approximation that is based on an asymptotic expansion of the logarithm of Bell polynomials. In this section, we will derive a formal expansion of $\log Y_N(1, 2!x_2, \dots, N!x_N)$ for fixed $\{x_2, x_3, \dots\}$. We are interested in the asymptotic behavior as $N \to +\infty$.

For simplicity, we will use following notations. The set $\{1, 2, \dots, K\}$ is denoted by $[K]$. The symbol $\mathcal{S}(X)$ (resp. $\mathcal{S}^2(X)$) stands for the set of non-empty families of disjoint subsets of X (resp. the set of non-empty families of disjoint subsets with cardinality ≥ 2 of X). The union $\cup_{I \in \mathcal{I}} I$ for $\mathcal{I} \in \mathcal{S}(X)$ is denoted by $\cup\mathcal{I}$. The set of those $\mathcal{I} \in \mathcal{S}(X)$ with $\cup\mathcal{I} = X$ is denoted by $\mathcal{S}^*(X)$. For $\mathcal{I}, \mathcal{J} \in \mathcal{S}(X)$ the relation $\mathcal{I} \ll \mathcal{J}$ means that each $I \in \mathcal{I}$ is included in some $J \in \mathcal{J}$ and that each $J \in \mathcal{J}$ includes at most one $I \in \mathcal{I}$. If $\cup\mathcal{I} = \cup\mathcal{J}$ ($= Y$, say), $\mathcal{I} \leq \mathcal{J}$ means the order relation that $\mathcal{I}$ is a finer partition of Y than $\mathcal{J}$. Let $\mu(\mathcal{I}, \mathcal{J})$ be the Möbius function for this order, that is,

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

If $\mathcal{I} \in \mathcal{S}(X)$, then let $\tau(\mathcal{I}) = \sum_{I \in \mathcal{I}} (\#I - 1)$, and $\zeta(\mathcal{I}) = \prod_{I \in \mathcal{I}} (-1)^{\#I - 1}(\#I - 1)$. For a sequence $t = \{t_i\}$, the sum $\sum_{i \in I} t_i$ over a subset I of indices is denoted by t_I. If $f(\boldsymbol{x}) = \sum_{\boldsymbol{a}} c_{\boldsymbol{a}} \boldsymbol{x}^{\boldsymbol{a}}$, then mindeg f is the minimum of $\{|\boldsymbol{a}|; c_{\boldsymbol{a}} \neq 0\}$. Also let $\xi(n) = (-1)^{n-1}(n-1)$.

PROPOSITION 5.1. *The following formal expansion holds*:

$$\log Y_N(1, 2!x_2, \dots, N!x_N) = \sum_{\boldsymbol{a} \in S_{N-1}} C_{\boldsymbol{a}}(N)(x_2, x_3, \dots, x_N)^{\boldsymbol{a}} \frac{1}{\boldsymbol{a}!}, \tag{5.1}$$

where $C_{\boldsymbol{a}}$ is a polynomial of the form;

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

PROOF. Let $\boldsymbol{x} = (x_1, \dots, x_N)$. From (2.3), (2.7) and (2.9) we have

$$\begin{aligned}
&(\partial/\partial \boldsymbol{x})^{\boldsymbol{a}} \log Y_N(1!x_1, 2!x_2, \dots, N!x_N)|_{\boldsymbol{x} = (1, 0, \dots, 0)} \\
&= \sum_{k=1}^{|\boldsymbol{a}|} (-1)^{k-1}(k-1)! B_{\boldsymbol{a},k} \left(\{ (N)_{\|\boldsymbol{b}\|}; \boldsymbol{b} \in S_N \} \right).
\end{aligned}$$

Therefore the following expansion follows:

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

where $\boldsymbol{y} = (x_1 - 1, x_2, \dots, x_n)$. If we let $x_1 = 1$ in this expansion, then only terms with $a_1 = 0$ remain and the right-hand side becomes as;

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

where $(0, \boldsymbol{a}) = (0, a_1, \dots, a_{N-1})$ and $\boldsymbol{z} = (x_2, \dots, x_N)$. Note that $|(0, \boldsymbol{a})| = |\boldsymbol{a}|$ and $\|(0, \boldsymbol{a})\| = \langle \boldsymbol{a}\rangle$. Also, from (2.5), we can see easily that

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

Thus the proposition has been proved.

PROPOSITION 5.2. *Polynomials $C_{\boldsymbol{a}}(y)$ are of degree $\|\boldsymbol{a}\| + 1$.*

The proof of this proposition will be given after providing two lemmas. Precisely, we will show that $\deg C_{\boldsymbol{a}} \le \|\boldsymbol{a}\| + 1$ first. That $\deg C_{\boldsymbol{a}} = \|\boldsymbol{a}\| + 1$ will be shown in Proposition 5.3.

LEMMA 5.1. *Define functions $F_{K,k}(\boldsymbol{t})$, $\boldsymbol{t} = (t_1, \dots, t_K)$, for $K \ge 1$ and $k \ge 1$ by*

$$F_{K,k}(\boldsymbol{t}) = \sum_{\mathcal{I}\in S^*([K])} (-1)^{\#\mathcal{I}-1}(\#\mathcal{I}-1)\left\{\prod_{I\in\mathcal{I}}(1+t_I)-1\right\}^k. \tag{5.2}$$

Then $\deg C_{\boldsymbol{a}} \le \|\boldsymbol{a}\| + 1$ *for all* $\boldsymbol{a} \in S_{N-1}$ *if* $\operatorname{mindeg} F_{K,k} \ge K + k - 1$ *for all* $K, k \ge 1$.

PROOF. From (2.8) and (2.9) we can show the relation

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

where $\boldsymbol{x} = (x_2, \dots, x_N)$. Using this relation, we have the generating function of $C_{\boldsymbol{a}}$ from (5.1);

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

Expand the left-hand side of the last relation directly and compare both sides, then

$$C_{\boldsymbol{a}}(y) = \sum_{n=1}^{|\boldsymbol{a}|}\frac{(-1)^{n-1}}{n}\sum_{\boldsymbol{b}_1+\cdots+\boldsymbol{b}_n=\boldsymbol{a}}\left\{\boldsymbol{a}!\Big/\prod_i \boldsymbol{b}_i!\right\}\left\{\prod_i (y)_{\langle \boldsymbol{b}_i\rangle}\right\},$$

where the innermost summation is taken over unordered $(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n)$. From the polynomial expansion formula, see Roman ((1984), Chapter 4),

$$\begin{aligned} C_{\boldsymbol{a}}(y) &= \sum_{k\geq 0} \Delta^k \{C_{\boldsymbol{a}}(0)\} \frac{(y)_k}{k!} \\ &= \sum_{k\geq 0} \left[\sum_{n=1}^{|\boldsymbol{a}|} \frac{(-1)^{n-1}}{n} \sum_{\boldsymbol{b}_1+\cdots+\boldsymbol{b}_n=\boldsymbol{a}} \left\{ \boldsymbol{a}! \Big/ \prod_i \boldsymbol{b}_i! \right\} \Delta^k \left\{ \prod_i (0)_{\langle \boldsymbol{b}_i \rangle} \right\} \right] \frac{(y)_k}{k!} \\ &\equiv \sum_{k\geq 0} C_{\boldsymbol{a},k} \frac{(y)_k}{k!}, \qquad \text{say,} \end{aligned}$$

where Δ is the forward difference operator $\Delta f(x) = f(x+1) - f(x)$ and $\Delta^k f(0) = \Delta^k f(x)|_{x=0}$.

It is convenient to associate a multi-index $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K) \in \{2, \ldots, N\}^K$, $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_K$, with each $\boldsymbol{a} \in S_{N-1}$ so that $K = |\boldsymbol{a}|$ and $\#\{i; \alpha_i = j+1\} = a_j$, $1 \leq j \leq N-1$. Note that $\langle \boldsymbol{a} \rangle = |\boldsymbol{\alpha}|$. For example, if $N = 5$ and $\boldsymbol{a} = (1,3,2,1)$, then $K = 7$ and $\boldsymbol{\alpha} = (2,3,3,3,4,4,5)$. Under this correspondence between $\boldsymbol{a}$ and $\boldsymbol{\alpha}$, each decomposition $\boldsymbol{b}_1 + \cdots + \boldsymbol{b}_n = \boldsymbol{a}$ of $\boldsymbol{a}$ corresponds to a partition $\mathcal{I} = \{I_1, \ldots, I_n\} \in \mathcal{S}^*([K])$ such that $\langle \boldsymbol{b}_i \rangle = \alpha_{I_i}$. Moreover it can be shown that

$$\sum \left\{ \boldsymbol{a}! \Big/ \prod_i \boldsymbol{c}_i! \right\} \prod_i (y)_{\langle \boldsymbol{c}_i \rangle} = n! \sum \prod_i (y)_{\alpha_{J_i}},$$

where the left-hand sum is taken over every different permutation $\{\boldsymbol{c}_i\}$ of $\{\boldsymbol{b}_i\}$ and the right-hand sum is taken over every partition $\mathcal{J} = \{J_1, \ldots, J_n\} \in \mathcal{S}^*([K])$ with $\{\alpha_j\}_{j\in J_i} = \{\alpha_j\}_{j\in I_i}$ for each i. Therefore, if we let, for $\boldsymbol{\alpha} \in S_K$,

$$C^*_{\boldsymbol{\alpha},k} = \sum_{\mathcal{I}\in\mathcal{S}^*([K])} (-1)^{\#\mathcal{I}-1} (\#\mathcal{I} - 1)! \Delta^k \left\{ \prod_{I\in\mathcal{I}} (0)_{\alpha_I} \right\}, \tag{5.3}$$

then $C^*_{\boldsymbol{\alpha},k} = C_{\boldsymbol{a},k}$ if $\boldsymbol{\alpha} \in \{2,3,\ldots,N\}^K$ corresponds to $\boldsymbol{a} \in S_{N-1}$. We should let $(x)_0 \equiv 1$.

Let $A = \prod_{I\in\mathcal{I}} (1 + t_I)$ for a fixed $\mathcal{I} \in \mathcal{S}^*([K])$. Then it can be shown that

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

Since $\Delta^k A^x = A^x (A-1)^k$,

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

From this relation, we can see easily that $F_{K,k}$ is the generating function of $\{C^*_{\boldsymbol{\alpha},k}\}$, that is,

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

Now the proof of Proposition 5.2 is finished if we can show $C_{\boldsymbol{a},k} = 0$ for $k > \|\boldsymbol{a}\|+1$. But, since $\|\boldsymbol{a}\| = |\boldsymbol{\alpha}| - K$, this results from $C^*_{\boldsymbol{\alpha},k} = 0$ for $|\boldsymbol{\alpha}| < K + k - 1$, which, in turn, is a result of $\operatorname{mindeg} F_{K,k} \geq K + k - 1$. Therefore Lemma 5.1 has proved.

LEMMA 5.2. *Functions $\{F_{K,k}\}$ satisfy the following recurrence relation for every K, $k \geq 1$:*

$$F_{K,k}(\boldsymbol{t}) - \left\{\prod_{i=1}^{K}(1+t_i) - 1\right\} F_{K,k-1}(\boldsymbol{t}) \tag{5.4}$$
$$= \sum_{\mathcal{I}\in\mathcal{S}^2([K])} \zeta(\mathcal{I})P_{\mathcal{I}}(\boldsymbol{t})$$
$$\times \left[\sum_{\mathcal{J}} \mu(\mathcal{I},\mathcal{J})F_{K-\tau(\mathcal{J}),k-1}\left(\{t_J\}_{J\in\mathcal{J}} \cup \{t_i\}_{i\notin\cup\mathcal{J}}\right)\right],$$

where the innermost summation is taken over those $\mathcal{J} \in \mathcal{S}^*(\cup\mathcal{I})$ with $\mathcal{I} \leq \mathcal{J}$, and

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

PROOF. We use notations like $F_{K,k}\left(\{t_i\}_{i\in I}\right)$ which have a natural meaning since $F_{K,k}$'s are symmetric functions. From the identity

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

for a fixed $L \subset [K]$, we have the following relation for each fixed $I \subset [K]$,

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

Hence for $\mathcal{I} \in \mathcal{S}^*(K)$

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

Using the last equality we can show the left-hand side of (5.4) is equal to

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

Let fix a $\mathcal{I} \in \mathcal{S}^2([K])$ and let $X = \cup\mathcal{I}$. For each $\mathcal{L} \in \mathcal{S}^*(X)$, let

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

Also let $\bar{H}(\mathcal{U}), \mathcal{U} \in \mathcal{S}^*(X)$, be the sum of $H(\mathcal{L})$ for $\mathcal{L} \in \mathcal{S}^*(X)$ with $\mathcal{U} \leq \mathcal{L}$. Then

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

From the Möbius inversion formula, see Comtet ((1974), Chapter 4, Supplement 15),

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

Therefore the assertion follows.

Proof of Proposition 5.2. Let us prove the assertion by induction on k for each $K \geq 1$. First note that $F_{1,k}(t_1) = t_1^k$. Therefore Proposition 5.2 is valid for $K = 1$ and $k \geq 1$. Also note that $\Delta\{(0)_n f(0)\} = 0$ if $n \geq 2$ and $= f(1)$ if $n = 1$. Hence, from (5.3), $C^*_{\boldsymbol{\alpha},1} = 0$ if $\boldsymbol{\alpha} \neq \mathbf{1} \equiv (1, 1, \ldots, 1)$ and $C^*_{\mathbf{1},1} = (-1)^{K-1}(K-1)!$. Therefore $F_{K,1} = (-1)^{K-1}(K-1)! \prod_{I \leq i \leq K} t_i$ and $\operatorname{mindeg} F_{K,1} = K$. Now assume that $K \geq 2$ and $\operatorname{mindeg} F_{K,k-1} \geq K + k - 2$. Then the mindeg of the second term of the left-hand side of (5.4) is $K + k - 1$. Also, the mindeg of each summand of the right-hand side of (5.4) is equal to

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

where we should note that $\tau(\mathcal{J}) = \#(\cup \mathcal{I}) - \#\mathcal{J}$. Therefore the induction is completed.

Proposition 5.3. *The exact form of* $C_{\boldsymbol{a},\|\boldsymbol{a}\|+1}$ *for* $\boldsymbol{a} \in S_{N-1}$ *is*

$$C_{\boldsymbol{a},\|\boldsymbol{a}\|+1} = (-1)^{|\boldsymbol{a}|-1}(\langle \boldsymbol{a} \rangle - 1)!(2, 3, \ldots, N)^{\boldsymbol{a}}. \tag{5.5}$$

For the proof of this proposition, we need the following lemma.

Lemma 5.3. *Define* $\kappa(n)$, $n \geq 2$, *by*

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

Then $\kappa(n) = (n-1)!$.

Proof of Lemma 5.3. The number of partitions $\mathcal{I}$ of $[n]$ with $\#\{I \in \mathcal{I}; \#I = i\} = x_i$, $1 \leq i \leq n$, is $n!/\{\prod_i x_i! \times \prod_i (i!)^{x_i}\}$, see Comtet ((1974), Chapter 5). Therefore

$$\begin{aligned} \kappa(n) &= \sum_{k=1}^{n} \sum \left[\left\{ n!(k-1)! \Big/ \prod_{i=2}^{n} x_i! \right\} \prod_{i=2}^{n} \left\{\frac{i-1}{i!}\right\}^{x_i}\right] \\ &= \sum_{k=1}^{n} (k-1)! B_{n,k}(0, 1, \ldots, n-k), \end{aligned} \tag{5.6}$$

where the innermost summation of the middle term is taken under conditions $\sum_{2\leq i\leq n} x_i = k$ and $\sum_{2\leq i\leq n} ix_i = n$. From (2.1) we have

$$\exp\left(u\left\{1+(t-1)e^t\right\}\right) = 1 + \sum_{n\geq 1}\left\{\sum_{k=1}^{n} u^k B_{n,k}(0,1,\dots,n-k)\right\}\frac{t^n}{n!}.$$

Multiply e^{-u} to both sides of this equality and integrate them over $(0,\infty)$, then

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

by (5.6). Therefore the assertion follows.

PROOF OF PROPOSITION 5.3. As we have seen, degrees of non-zero terms of $F_{K,k}$ are at least $K+k-1$. Let $G_{K,k}$ be the sum of those terms of $F_{K,k}$ with degree $K+k-1$. Then, extracting terms with degree $K+k-1$ from both sides of (2.2), we can derive the recurrence relation;

$$\begin{aligned}(5.7)\quad G_{K,k}(t) = &\left\{\sum_{i=1}^{K} t_i\right\} G_{K,k-1}(t)\\ &+\sum_{n=2}^{K}(-1)^{n-1}\kappa(n)\\ &\times\left[\sum_{I\in[K],\#I=n}\left\{\prod_{i\in I} t_i\right\} G_{K+1-n,k-1}\left(\{t_I\}\cup\{t_i\}_{i\notin I}\right)\right].\end{aligned}$$

Now we will prove by induction on both K and k that

$$(5.8)\qquad G_{K,k}(t) = (-1)^{n-1}\phi_K(k)\left\{\prod_{i=1}^{K} t_i\right\}\left\{\sum_{i=1}^{K} t_i\right\}^{k-1}$$

for all $K, k \geq 1$, where $\phi_K(k)$ are constants determined later. From the proof of Proposition 5.2, $G_{K,1}$ for all $K\geq 1$ has the form (5.8) with $\phi_K(1) = (K-1)!$. If (5.8) is valid for $k = m-1$ and $K = 1,2,\dots,M-1$, then, by (5.7), it follows in fact after some manipulations that

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

where we have set

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

Again, we can show by induction, using Lemma 5.3 and the last recurrence relation, that $\phi_K(k) = (K+k-2)!/(k-1)!$. Thus we have proved that $G_{K,k}$ is equal to

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

that is, for $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K) \in \{1,2,\ldots\}^K$ with $|\boldsymbol{\alpha}| = K+k-1$,

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

But, from the correspondence $\boldsymbol{\alpha} \in \{2,\ldots,N\}^K \to \boldsymbol{a} \in S_{K-1}$ and $C^*_{\boldsymbol{\alpha},k} = C_{\boldsymbol{a},k}$, the last relation implies

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

for $\boldsymbol{a} \in S_{N-1}$, which is the assertion.

Now we can give approximations of r_n based on (5.1) and (5.5). Let $x_i = (-1)^{i-1}P_i/i$ in (5.1) and approximate $C_{\boldsymbol{a}}(n)$ by its highest order term $C_{\boldsymbol{a},||\boldsymbol{a}||+1}(n)_{||\boldsymbol{a}||+1}/(||\boldsymbol{a}||+1)!$. Then we get

$$(5.9)\quad \frac{1}{n}\log Y_n(1, -P_2, \ldots, (-1)^{i-1}(i-1)!P_i, \ldots, (-1)^{n-1}(n-1)!P_n)$$
$$\simeq \sum_{\boldsymbol{a}\in S_{n-1}} (-1)^{\langle\boldsymbol{a}\rangle-1}(\langle\boldsymbol{a}\rangle-1)! \frac{(n-1)_{||\boldsymbol{a}||}}{(||\boldsymbol{a}||+1)!} \frac{(P_2,\ldots,P_n)^{\boldsymbol{a}}}{\boldsymbol{a}!}.$$

Note that $(n-1)_{||\boldsymbol{a}||} \simeq n^{||\boldsymbol{a}||}$ as $n \to \infty$ and $(P_2,\ldots,P_n)^{\boldsymbol{a}} = O(c^{||\boldsymbol{a}||})$, and, hence, summands in the right-hand side of (5.9) with large $||\boldsymbol{a}||$ are negligible as far as cn remains small. Selecting appropriate terms, we have following approximations to r_n:

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

The approximant $\rho_{n,1}$ with $P_2 = 1/365$ is well-known in the birthday problem, see Feller (1968). Actually, in the birthday problem, maximas of $|r_n - \rho_{n,m}|$ for

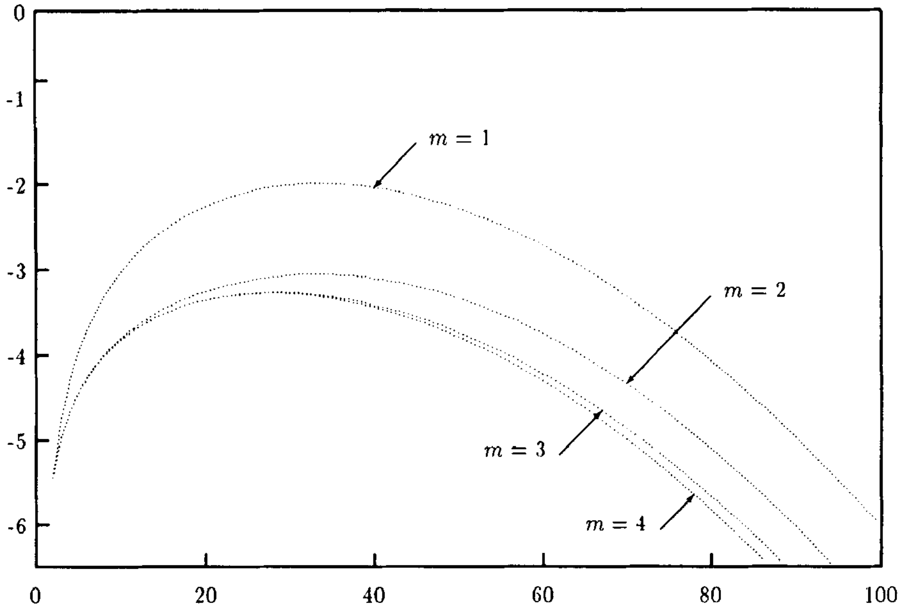

Fig. 1. Birthday problem: approximation of r_n by $\rho_{n,m}$. Curves of $\log_{10}|r_n - \rho_{n,m}|$ for $m = 1, 2, 3, 4$.

$5 \leq n \leq 100$, $m = 1, 2, 3, 4$, are about 0.01034684, 0.00090218, 0.00055420 and 0.00054270 respectively, showing excellent agreements, see Fig. 1.

Remark. Note that the approximation (5.9) is formal. It seems very difficult to give a rigorous error bound. Even if there were such a bound, it would be probably effective only asymptotically as $n \to \infty$ while $P_a \to 0$. If so, it would not be of use in our case where we are interested in fairly small n. Arratia *et al.* (1989) showed that the approximant $\rho_{n,1}$ results from the Chen-Stein Poisson approximation of the distribution of a sum of Bernoulli random variables and gave an exact bound for approximation error $|r_n - \rho_{n,1}|$. However their bound has a meaning only if n is large and, on the contrary, $\binom{n}{2}P_2$ remains moderate in size. For example, their bound for the birthday problem is less than 0.01 only for $n \leq 9$.

6. Surname problem

We now apply the preceding results to the *real* surname problem. To this end, we must know the distribution of surnames in Japan. But this is a difficult problem. There are extremely many surnames in Japan and no one knows exactly how many kinds there are. Niwa (1978) records 110,867 different surnames. Niwa, a researcher of Japanese surnames, has collected about 120,000 kinds, see Niwa (1980). There are many reasons for this abundance, which are historical, cultural and linguistical. Also there are only two surveys of the global distribution of surnames in Japan. The first such study was done by the Univac Japan Co. in order to computerize the handling of names by "kanji", see Tanaka (1972). In

Japan an ideogram of Chinese origin, called *kanji*, and two types of phonograms, called *kana*, are used simultaneously. Names are fundamentally denoted by kanji. By the way, this is one of main reasons why there are so many surnames. Names that are identical phonologically, therefore have the same kana notation, may have completely different kanji notations.

Sources of two existing surveys were computer record files of insurants of life insurance companies. Among them, that of Daiiti Life Insurance Co. is the biggest in data size and we will use it in the sequel, see Daiiti Life Insurance Co. (1987). It is based on about 8,320,000 insurants and their 11,098,833 insurances in 1986, when Japanese total population was about 121,000,000. However the published data are only percentages of surnames till the 200th rank, see Table 1.

Table 1. Typical Japanese surnames up to 21th rank: a part of Daiiti Life Insurance data.

Sato 1.583%	Suzuki 1.332%	Takahashi 1.132%	Tanaka 1.061%	Watanabe 1.007%	Ito 0.950%	Nakamura 0.864%
Yamamoto 0.856%	Kobayashi 0.812%	Saito 0.799%	Kato 0.720%	Yoshida 0.670%	Yamada 0.661%	Sasaki 0.590%
Matsumoto 0.527%	Yamaguchi 0.519%	Kimura 0.476%	Inoue 0.470%	Abe 0.464%	Hayashi 0.425%	Shimizu 0.412%

We can point out several problems about this data. Clearly this is not data of random sampling and, moreover, actual percentage numbers is not calculated by number of insurants but by number of insurances, so there are many duplications. Probably these two points do not cause a serious bias. Also the distribution of surnames varies locally, but the local distribution of insurants is roughly proportional to actual local populations. The most annoying feature of this data is the fact that computer files record insurant names by kana, that is, phonologically because computers could not handle kanji characters for a long time. For example, "Ito", the surname of famous probabilist K. Ito, has the 6th rank. But actually there are two major Ito's which have different kanji notations and have different origins and, in addition, there are more than 16 rare Ito's with different kanji notations.

But, since we have no alternative, we will simply neglect this point and assume a family of names with the same kana notation as a unit. We show in Table 2 cumulative percentages according to ranks which is a part of the Univac study, see Tanaka (1972). This was calculated by computer files of Daihyaku Life Insurance Co. It is smaller in data size, based on 715,815 insurances, but we see immediately that the distribution has an exceptionally long tail even if we do not take too rare names into account. Therefore it seems hopeless to fit ordinary distributions to this data. From some experiments we knew that a function of the form $f(n) = d/n^a$, where a is about 0.7, shows a good fitting at least within the range of ranks ≤ 200. If $a > 1$, this is the zeta, or Zipf, distribution if normalized. On this account it may be suggestive to note that the Zipf distribution can be the distribution of rank-frequency of sizes approximately as shown in Hill (1974). But, if $a \leq 1$ this is a divergent series and in order to get a convergent series we modify this function by multiplying a convergence factor c^n with $c < 1$. The convergence factor c must

Table 2. Cumulative distributions of Japanese surnames: Daihyaku Life Insurance data A (%) and estimated distributions B (%).

rank	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

be near to 1 in order to have a long-tailed distribution. As a result we chose the function

$$f(n : a, b, c, d) = d\frac{c^n}{(n+b)^a}, \qquad n = 1, 2, 3, \dots . \tag{6.1}$$

The function defined by the sum $\sum_{n=0}^{\infty} f(n : a, b, c, 1)$ is equal to $\Phi(c, a, b)$ where Φ is the generalized (Riemann's) zeta function, see Gradshteyn and Ryzhik (1980). Finally, since it is difficult to compute the infinite sum $\Phi(c, a, b)$ numerically we truncate the summation range as

$$p(n : a, b, c) = \xi(a, b, c)^{-1}\frac{c^n}{(n+b)^a}, \qquad 1 \leq n \leq 120{,}000, \tag{6.2}$$

where $\xi(a, b, c)$ is the normalized constant

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

Actually, bias caused by truncation is seen to be negligible. Next we tried to fit the function by a non-linear regression method. A straightforward fitting based on the three-parameter probability function (6.2) failed because of the inaccuracy of calculated values and the long computation time of ξ. If we fit the four-parameter function (6.1), we get values $a = 0.9474$, $b = 5.798$, $c = 1.002$ and $d = 0.09464$. Although these parameters give a fairly good fitting with the coefficient of determination $R^2 = 99.38\%$, the value c which is larger than 1 makes the extrapolation beyond $n > 200$ absurd. As a result, we took the strategy to fit parameters (a, b, d) of the function (6.1) for each trial value $c < 1$ and to choose c so that the sum of resulting function (6.1) for $1 \leq n \leq 120{,}000$ is as near to 1 as possible. Thus we obtained values $a = 0.7570$, $b = 3.648$ and $c = 0.9996$. Finally we substituted these values into the function (6.2), $\xi(a, b, c) = 19.80427$, and estimated the distribution of Japanese surnames. The coefficient of determination is $R^2 = 99.27\%$. If we fit the same function to the Daihyaku Life Insurance data, the coefficient of determination is $R^2 = 97.98\%$. We show estimated cumulative distributions in Table 2. By the way, the calculation of ξ and power sums P_a, in fact all computations except non-linear regressions, were done using UBASIC, a BASIC interpreter on MS-DOS based personal computers, developed by Professor U. Kida of Rikkyo University which is capable of making arithmetics of 2,600 digits fixed point numbers (8th version).

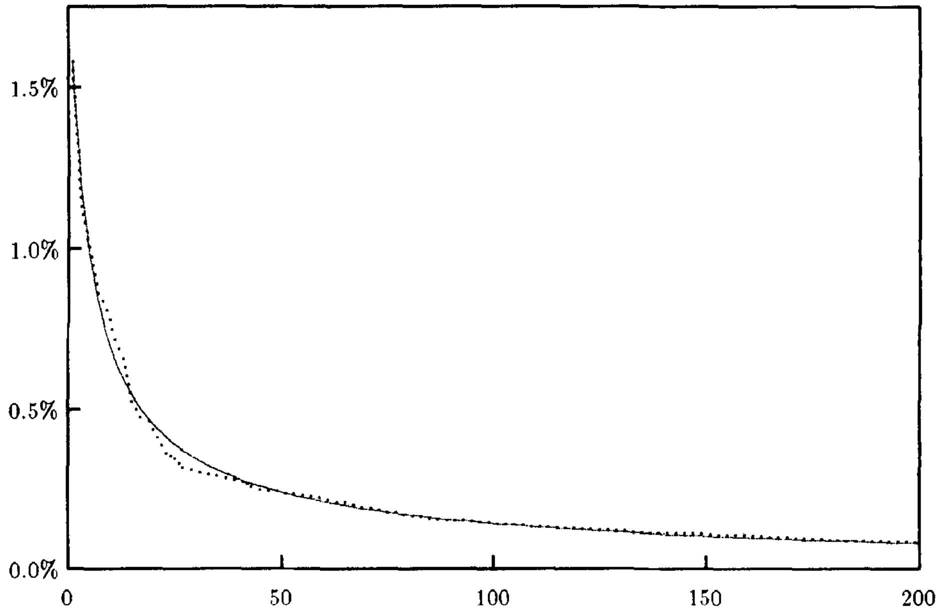

Fig. 2. Distribution of surnames up to 200th rank and fitted curve.

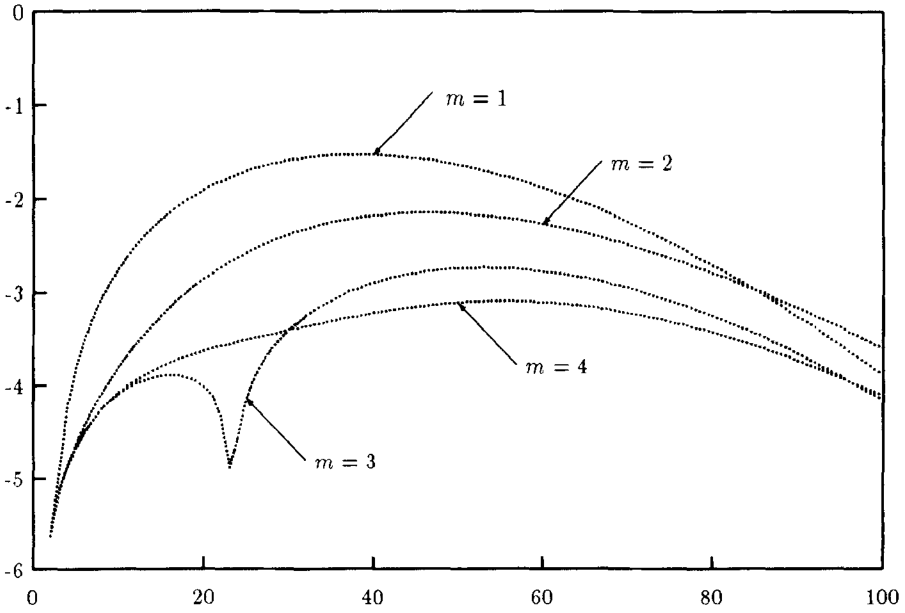

Fig. 3. Surname problem, approximation of r_n by $\rho_{n,m}$. Curves of $\log_{10} |r_n - \rho_{n,m}|$ for $m = 1, 2, 3, 4$.

Now we can calculate the surname probabilities R_n. First we use the recurrence relation (3.2) directly. The result is tabulated in Table 3, see also Fig. 2. The coincidence probability exceeds 50% at $n = 27$, $R_{27} = 51.153\%$. Also typical cases are $R_{18} = 27.327\%$, $R_{38} = 75.332\%$, $R_{41} = 80.250\%$, $R_{50} = 90.727\%$, $R_{57} = 95.289\%$ and $R_{71} = 99.028\%$. In Fig. 3 the degree of approximations to the surname probability r_n by approximants $\rho_{n,m}$, $m = 1, 2, 3, 4$, is shown. The cusp in the curve of $\rho_{n,3}$ is due to the fact that r_3 and $\rho_{n,3}$ happen to be very close at that point.

Table 3. Probabilities (%) of coincidence of surnames.

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

7. True birthday problem

As the second application we evaluate the *true* birthday coincidence probabilities in Japan. The true distribution of birthdays shows a considerable seasonal variation. Also it varies with generations. For example, the maximal monthly birth number was 154,114 (January) and the minimal was 85,469 (June) in 1900. The proportion is 180:100. On the other hand, the maximal monthly birth number was 134,734 (August) and the minimal was 116,152 (February) in 1980. The proportion is 116:100. Also the annual birth number varies from generation to generation. We dare not to refer to false registrations of birthdays due to preference or avoidance of certain calendar days. Klotz (1979) believed that physician's convenience is an important factor.

The author could not get daily birth number data and had to estimate them from partial data taken from existing literature. The basic data of the subsequent work are monthly birth numbers for 1900(5)1940,1947,1950(5)1980, and 1982(1)1987, see Ministry of Health and Welfare (1988), and populations in 1988 of Japanese born in 1900(1)1988, see Management and Coordination Agency (1990). We use a simple piecewise linear interpolation. Let $M(i)$, $1 \leq i \leq 12$, be monthly birth numbers of a year. We divide $M(i)$ by both the annual birth number and the number of days of the month i and get the ratio $m(i)$. Let $d(i)$ be the middle day of the month i. Next we make a piecewise linear interpolation of points $P(i) = (d(i), m(i))$, $0 \leq i \leq 13$, where points are arranged cyclically so that $P(0) = (d(12) - 365, m(12))$, $P(13) = (d(1) + 365, m(1))$, and, after normalization, get daily birth rates $n(j)$, $1 \leq j \leq 365$. We neglect the 29th of February of leap years. As to daily birth rates of those years for which monthly birth numbers are not available, we make the linear interpolation between corresponding daily birth rates of two adjacent years for which we know monthly birth numbers. Finally, we multiply living population of each generation in 1988 to each estimated daily birth rate for each generation born in 1900(1)1987.

Thus, obtained daily birth rates for total Japanese population show also a considerable fluctuation. The maximal rate is 0.343% (January 16) and the minimal is 0.236% (June 15). Recall that 1/365=0.274%. Their proportions are 125:100:86. Now we can evaluate the true birthday probabilities. Although birth rates are far from uniform, we find that corresponding birthday probabilities are almost unchanged. The maximal discrepancy is 0.370% at $n = 27$. Klotz (1979) observed the same phenomenon. By the way, it is known that birthday probabilities are smallest for the uniform birth rate case, see Bloom (1973) and Munford (1977). Munford stated that this fact explains why in practice coincidental birthdays are observed more frequently than the uniform theory predicts. But our example and Klotz's example contradict to his belief. Presumably Munford generalized his personal observation too hastily.

This stableness of coincidence probabilities may be explained as follows. From the previous discussions, we see that r_n can be already approximated fairly well by $\rho_{n,1}$ which depends only on P_2. Let $p_i = (1 + \epsilon_i)/365$. Then $\sum_i \epsilon_i = 0$ and

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

That is, a fluctuation of birthday rates from 1/365 results in the change of the value P_2 from 1/365 to $(1 + \sigma^2)/365$ where σ^2 is the variance of $\{\epsilon_i\}$. Hence the relative difference between two values of $\rho_{n,1}$ is approximately equal to

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

In our example $\sigma^2/365$ is about 0.00003, which explains why birthday probabilities are almost unchanged. We could use more sophisticated interpolation techniques to estimate daily birth rates. Nevertheless our result suggests that the true birthday probabilities are not so different from idealized ones.

Acknowledgements

The author wants to express his thanks to Professor M. Sibuya for his kind encouragement and bringing the paper of Flajolet *et al.* to his notice, and to Professor M. Maejima who kindly taught him the paper of Klotz. Also the author owes numerous valuable suggestions and corrections to referees.

References

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

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

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

Bolotonikov, Yu. V. (1968). Limiting processes in a model of distribution of particles into cells with unequal probabilities, *Theory Probab. 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 Probab. Appl.*, **10**, 149–154.

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

Daiiti Life Insurance Co. (ed.) (1987). *Surnames and Names*, Kouyuu Publishing Co., Tokyo (in Japanese).

Fang, K.-T. (1985). Occupancy problems, *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., Gardy, D. and Thimonier, L. (1988). Probabilistic languages and random allocations, *Lecture Notes in Comput. Sci.*, **317**, 239–253, Springer, Berlin.

Flajolet, P., Gardy, D. and Thimonier, L. (1991). Birthday paradox, coupon collectors, caching algorithms and self-organizing search, *Discrete Appl. Math.* (to appear).

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.

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

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

Kolchin V. F., Sevast'yanov, B. A. and Chistyakov, V. H. (1978). *Random Allocation* (translation ed. A. V. Barakrishna), Wistons and sons, Washington D.C.

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

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

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, *Amer. Statist.*, **31**, 119.

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

Niwa, M. (ed.) (1978). *Japanese Surnames*, Vol. 1 and 2, Nippon Keizai Shinbun-Sha Co., Tokyo (in Japanese).

Niwa, M. (1980). *Origins of Surnames*, Kadokawa Book Co., Tokyo (in Japanese).

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

Tanaka, K. (1972). Statistics of Japanese surnames and names, *Gengo-Seikatu*, **254**, 72–79 (in Japanese).