# Partition of homonyms by accents and a generalized family of Stirling distributions

Masaaki Sibuya, Takachiho University and Keio University

sibuyam@1986.jukuin.keio.ac.jp

Cherry Bud Workshop 2006

**Abstract**

Homonyms are partitioned into smaller groups by their accents. The property of accents in English, Japanese and Chinese languages is different, but has a common role to distinguish homonyms. The difference of the role of accents in the three languages is shown by a parameter of the generalized Stirling distribution, derived from Pitman's random partition, fitted to dictionary datasets.

## 1 Introduction and preliminaries

Homonyms, or homographs more exactly, are words of the same spelling, appearing as different items in a dictionary. The number $y_{n,k}$ of homonym groups of $n$ words clustered into $k$ types of accents, $k = 1, \ldots, n, n > 1$, are collected from an English, a Japanese and a Chinese dictionaries, by Sibata and Shibata (1990), to study the role of accents to distinguish homonyms in the three languages. To these datasets, a sort of the Stirling family of distributions, derived from Pitman's random partition is fitted. Values of a parameter distinguish clearly the three languages.

Stirling numbers have been extended in many directions. Hsu and Shiue (1998) unified them through a factorial product representation of the polynomial ring. In this paper, the Stirling family of discrete probability distributions, Sibuya (1988) and Nishimura and Sibuya (1997), is extended along with the new unification.

## Random partition

Suppose a set of $n$ elements ($n \in \mathcal{N}_+$) is partitioned randomly into $k$ nonempty and mutually exclusive subsets, or clusters, of size $c_1, \ldots, c_k$ ($c_i > 0,\ 1 \le i \le k,\ \sum_{i=1}^{k} c_i = n$). If the elements are undistinguishable, and the order of clusters are irrelevant, the available statistics are the ordered size of clusters and summarized by $S_j = \sum_{i=1}^{k} I[c_i = j],\ j = 1, \ldots, n$. The quantity $S = (S_1, \ldots, S_n)$ is a random partition of an integer $n$, such that $S_j \ge 0$ and $\sum_{j=1}^{n} jS_j = n$, and called *size index* (frequency spectrum, or frequency of frequencies). A celebrated distribution of size index is *Pitman's random partition* (or Pitman's sampling formula), defined by

$$p(\boldsymbol{s}; n, \theta, \alpha) = \frac{n!\theta(\theta+\alpha)\ldots(\theta+(k-1)\alpha)}{\theta^{\overline{n}}} \prod_{j=1}^{n} \left( \frac{(1-\alpha)^{\overline{j-1}}}{j!} \right)^{s_j} \frac{1}{s_j!}, \tag{1}$$

$$\boldsymbol{s} = (s_1, \ldots, s_n) \in \mathcal{C}_n, \quad k = \sum_{j=1}^{n} s_j, \quad n \in \mathcal{N},$$

where $\mathcal{C}_n = \{\boldsymbol{s} = (s_1, \ldots, s_n),\ s_j \ge 0,\ j = 1, \ldots, n;\ \sum_{j=1}^{n} js_j = n\}$ and $y^{\overline{j}} = y(y+1)\ldots(y+j-1)$, while $y^{\underline{j}}$ is the descending factorial product to be used later. The expression (1) is positive for any $1 \le k \le n$ in $\{(\alpha, \theta) :\ 0 \le \alpha < 1,\ -\alpha < \theta \quad \text{or} \quad \theta = -m\alpha > 0,\ m = 1, 2, \ldots\}$. We are concerned with, in this paper, only the case $\theta = m\gamma,\ \gamma = -\alpha > 0,\ m = 2, 3, \ldots$

Pitman's random partition (1) of the case $\gamma = -\alpha > 0$ can be constructed in another way. Suppose that $\boldsymbol{X} = (X_1, \ldots, X_m)$ follows the symmetric(homogeneous) multivariate negative hypergeometric distribution, $\mathsf{MvNgHg}(m, n; (\gamma, \ldots, \gamma))$,

$$P\{\boldsymbol{X} = \boldsymbol{x}\} = \left( \prod_{j=1}^{m} \frac{\Gamma(\gamma/m + x_j)}{\Gamma(\gamma/m)x_j!} \right) \Big/ \frac{\Gamma(n+\gamma)}{\Gamma(\gamma)n!}, \tag{2}$$

$$x_j \ge 0,\ j = 1, \ldots, m,\ \sum_{j=1}^{m} x_j = n; \quad \gamma > 0.$$

Put $S_j = \sum_{i=1}^{m} I[X_i = j],\ j = 1, \ldots, n$, and $S = (S_1, \ldots, S_n)$, is Pitman's random partition, (1). This genesis gives the intuitive meaning of the parameter $\gamma$.

For Pitman's random partition, see, e.g., Johnson, Kotz and Balakrishnan (1997), Pitman (1999, 2002), Sibuya and Yamato (2000), Yamato, Sibuya and Nomachi (2001). Hoshino (2001) applied the random partition to another research field.

## Generalized Stirling Distribution

If $S = (S_1, \ldots, S_n)$ is Pitman's random partition (1), the number of clusters $K = \sum_{j=1}^{n} S_j$ has the probability mass function (pmf),

$$P\{K = k\} = \frac{m^{\underline{k}}\gamma^k}{(m\gamma)^{\overline{n}}} s_\gamma(n,k), \quad 1 \le k \le \ell; \quad \ell = \min(n,m), n, m \in \mathcal{N}, \tag{3}$$

where $s_\gamma(n,k)$ is a type of the generalized Stirling numbers, defined by the polynomial identity

$$x^{\overline{n}} = \sum_{k=1}^{n} s_\gamma(n,k) \prod_{j=0}^{k-1}(x - j\gamma). \tag{4}$$

Actually $s_\alpha(n,k)$ is a polynomial in $\gamma$ of degree $n-k$ with integer coefficients. For the generalized Stirling number see Pitman (2002) and the Appendix. The distribution (3) will be called *Generalized Stirling Distribution* and denoted by $\mathsf{GStr}(n,m,\gamma)$. The shapes of the pmf (3) are shown in **Figs. 1 and 2**. If $g$ or $m$ is small the pmf is decreasing, if $g$ or $m$ is larger the pmf is increasing, otherwise it is unimodal.

Let $p(k;m,m,\gamma)$ denote the pmf (2) of $\mathsf{GStr}(n,m,\gamma)$. It satisfies the recurrence formula

$$p(k;n+1,m,\gamma) = \frac{1}{m\gamma+n}\left((k\gamma+n)\; p(k;n,m,\gamma) + (m-k+1)\;\gamma\; p(k-1;n,m,\gamma)\right), \tag{5}$$

with $p(1;1,1,\gamma) = 1$, and

$$p(1;n,m,\gamma) = \frac{(1+\gamma)^{\overline{n-1}}}{(1+m\gamma)^{\overline{n-1}}}; \quad n,m > 1,$$

$$p(n;n,m,\gamma) = \frac{m^{\underline{n}}\gamma^n}{(m\gamma)^{\overline{n}}} \begin{cases} > 0, & m \ge n, \\ = 0, & m < n. \end{cases}$$

The recurrence formula is used for the numerical computation of the pmf. Moreover, using the formula

$$\mathrm{E}(m-K)^{\underline{r}}) = m^{\underline{r}}\frac{((m-r)\gamma)^{\overline{n}}}{(m\gamma)^{\overline{n}}}, \quad r = 1,2,\ldots$$

In particular

$$\mathrm{E}(K) = m\left(1 - \frac{((m-1)\gamma)^{\overline{n}}}{(m\gamma)^{\overline{n}}}\right) =: \mu, \tag{6}$$

and

$$\mathrm{Var}(K) = \frac{m^{\underline{2}}((m-2)\gamma)^{\overline{n}}}{(m\gamma)^{\overline{n}}} - (m-\mu)\;(m-1-\mu).$$

Pitman's random partition, as well as $\mathsf{GStr}(n,m,\gamma)$, is generated by an urn model. There are $m$ urns, $\mathrm{U}_1, \ldots, \mathrm{U}_m$, and balls are randomly thrown into $\mathrm{U}_1, \mathrm{U}_2, \ldots$, sequentially as

follows. The first one is put into $\mathrm{U}_1$. If $\mathrm{U}_1, \ldots, \mathrm{U}_k$ are occupied by $c_1, \ldots, c_k$ balls respectively ($c_i > 0, i = 1, \ldots, k,\ k < m;\ \sum_{i=1}^{k} c_i = n$), the $n+1$ st ball is thrown into

$$\begin{array}{lll} \text{urn } \mathrm{U}_i \ \text{(occupied)} & \text{with the probability} & \frac{c_i+\gamma}{m\gamma+n}, \quad i = 1, \ldots, k, \\ \text{urn } \mathrm{U}_{k+1} \ \text{(empty)} & \text{with the probability} & \frac{(m-k)\gamma}{m\gamma+n}. \end{array}$$

At the $n$-th stage of this process, the size index follows (1) and the number $K$ of occupied urns follows $\mathsf{GStr}(n, m, \gamma)$. In this model any two balls are put into different urns with the probability

$$P\{K = 2|n = 1\} = \frac{(m-1)\gamma}{m\gamma+1} =: \rho \in (0, 1 - 1/m), \tag{7}$$

and in the same urn with the probability

$$P\{K = 1|n = 1\} = \frac{1+\gamma}{m\gamma+1} = 1 - \rho.$$

Conversely $\gamma = \gamma(\rho; m) = \rho/(m(1-\rho)-1)$. $\rho = \rho(\gamma; m),\ m = 2, 3, \ldots$ is increasing, and concave in $\gamma$ and the upper limit is $(m-1)/m$ ($\gamma \to \infty$).

# 2 Homonyms

## English

*Homonym* is "any of two or more words spelt and pronounced alike but different in meaning" (Longman Dictionary of the English Language). More precisely, homonym is *homophone* (has the same sound) or *homograph* (written the same). For example the word "desert" has three entries in the dictionary:

$$\begin{array}{ll} \text{desert}^1: & \text{to leave without permission,} \\ \text{desert}^2: & \text{an arid region,} \\ \text{desert}^3: & \text{the fact of deserving reward or punishment,} \end{array}$$

and these are homographs, partitioned into two accent groups $\{\text{desert}^1,\ \text{desert}^3\}$ and $\{\text{desert}^2\}$. While,

$$\text{dessert}: \quad \text{a sweet course,}$$

is a homophone of $\{\text{desert}^1, \text{desert}^3\}$.

Homographic words are distinguished sometimes by accents. In this paper, *homonym is defined as a group of homographic words.* That is, 'desert' is a homonym group of three

words (or size three), separated into two clusters by accents. The number of accent clusters are less than or equal to the number of syllables and the size of the group.

About two-thirds of English homonym groups are one-syllable words, and they are out of scope of this paper. It is remarkable that homophones are almost all one-syllable words.

## Japanese

In written Japanese language, two types of characters, Kanji (ideogrammatic Chinese character) and Kana (phonogram), are used. The reading of a Kanji is not unique, and we are concerned here with only Kana spelling. Kana characters do not correspond to syllables exactly, and linguists use the notion of 'mora' as a unit of pronunciation. Accents are intonation of moras. In this paper, a Japanese homonym group is the collection of the same Kana spelling, and homographs are separated into clusters by mora-intonation. For example,

hashi: end, HAshi: chopstick, haSHI: bridge.

Since there is a word of flat tone moras, a homonym group of $r$-mora words can have $r + 1$ accent clusters.

## Chinese

In written Chinese, a character corresponds to a syllable, which is expressed by phonetic 'pinyin' spelling, and may have four tones. For example,

| | |
|---|---|
| mā : mother, wipe, pair | má : hemp, numb |
| mǎ : horse, | mà : swear |

In this paper, a homonym group is defined as that of words of the same pinyin spelling, disregarding the four tones. In this example, 'ma' is a homonym group of seven words partitioned into four accent clusters. A group of $r$ characters can have logically $4^r$ accent clusters, but actually there are not so many kinds of accent. Words in a accent cluster have different meaning and different ideogrammatic characters. (In China, Main Land, words of the same pronunciation tend to be written by the same character. That is, Chinese characters are becoming more phonetic.)

## Homonym datasets

Shibata and Shibata (1990) surveyed all the entries of three small-size dictionaries of English, Japanese and Chinese languages to count homonyms. It is evident that the role of accents is quite different in these languages, and the authors intended to find a numeric measure

to show the difference. There are 1430 English homonym groups and 918 of them are one-syllable words. There are 3038 Japanese homonym groups, and 5019 Chinese groups. The difference of these numbers is remarkable, but the frequency aspects of homonyms are not analyzed in this paper.

The data shown in Shibata and Shibata (1990) are, for each of the three languages, $y_{nmk}$, the number of homonym groups with $n$ homographic words (size $n$), which can have $m$ different accents (possible number of accent clusters) and actual number of clusters $k$, $1 \le k \le \ell,\ \ell = \min(n, m)$.

For English words of $r$ syllables, $m = r$; for Japanese words of $r$ moras, $m = r + 1$; and for Chinese words of $r$ characters ($r$ pinyin syllables), $m = 4^r$. The values of $y_{nmk}$ for the three languages are shown in Tables 6–**??** in Appendix.

The dataset $(y_{nm1}, y_{nm2}, \dots)$ is assumed to follow the multinomial distribution $\mathsf{Mn}(y_{nm\cdot}, (p_1, \dots, p_k))$, given $y_{nm\cdot} = \sum_{k=1}^{\ell} y_{nmk}$, where $p_k = p(k; m, n, \gamma_{nm})$ is the $\mathsf{GStr}(n, m, \gamma_{nm})$ probabilities.

Shibata and Shibata (1990) fitted a modification of binomial distribution to $y_{nmk}$. Sibuya (1991) disregarded $m$, regarding syllables to be formed randomly, and fitted Ewen's random partition to $y_{nm\cdot}$, to find a surprisingly good fit.

# 3 Inference

The likelihood $p(k; n, m, \gamma)$ behaves unfavorably for smaller and larger values of $k$. It is shown that $p(1; n, m, \gamma) \to 1$ ($\gamma \to 0$), and $p(n; n, m, \gamma) \to 1$ ($\gamma \to \infty$). Hence the maximum likelihood method cannot be used. The expectation $\mu$, (6), of $K$ satisfies

$$\mu \ < \ m\left(1 - \left(1 - \frac{1}{m}\right)^n\right), \quad \gamma \in \mathcal{R}_+$$

and it is a strictly decreasing and convex function of $\gamma$, and the moment method is applicable when $K$ is not very large. **Tables 1 and 2** show the moment estimates of $\gamma_{nm}$ and corresponding $\rho_{nm}$ calculated by (7) for the three languages. The values give rough ideas of the datasets, and entries for $l = \min(n, m) = 2$, and $\widehat{\gamma} = 0$ or Inf, are also minimum chi-square estimates. Hence the minimum chi-square method is used further for estimating $\gamma$. The result for the Japanese datasets is shown in **Table 3**, and the $p$-values of chi-square statistics are shown by the PP-plot in **Fig. 3** because the degrees of freedom are not the same. Since the observed frequency is small in some entries, the PP-plot is not against the proposed mode.

English

| $m$ | $n=2$ | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| 2 | 0.026 | 0.063 | 0 | NA | NA |
| 3 | 0.009 | 0.000 | NA | NA | NA |
| 4 | 0.000 | NA | NA | NA | NA |
| 5 | 0.000 | NA | NA | NA | NA |

Japanese

| $m$ | $n=2$ | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| 2 | 6.000 | 0.400 | 1.084 | Inf | Inf | Inf | Inf |
| 3 | 0.529 | 0.730 | 0.343 | 0.281 | 0.331 | 0.453 | 0.000 |
| 4 | 0.162 | 0.142 | 0.122 | 0.193 | 0.189 | 0.142 | 0.081 |
| 5 | 0.045 | 0.047 | 0.049 | 0.060 | 0.078 | 0.000 | 0.031 |
| 6 | 0.026 | NA | NA | NA | NA | NA | NA |
| 7 | 0.059 | NA | NA | NA | NA | NA | NA |
| 8 | 0.000 | NA | NA | NA | NA | NA | NA |

Chinese

| $m$ | $n=2$ | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| 4 | Inf | Inf | 51.641 | 4.736 | 374.776 | 2.827 | Inf |
| $4^2$ | 0.502 | 0.767 | 0.716 | 0.794 | 0.956 | 1.834 | 0.804 |
| $4^3$ | 0.079 | Inf | NA | NA | NA | NA | NA |
| $4^4$ | 0.020 | NA | NA | NA | NA | NA | NA |

Table 1: Moment estimates $\widehat{\gamma}_{nm}$.

English

| $m$ | $n=2$ | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| 2 | 0.025 | 0.056 | 0 | NA | NA |
| 3 | 0.017 | 0.000 | NA | NA | NA |
| 4 | 0.000 | NA | NA | NA | NA |
| 5 | 0.000 | NA | NA | NA | NA |

Japanese

| $m$ | $n=2$ | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| 2 | 0.462 | 0.222 | 0.342 | 1.000 | 1.000 | 1.000 | 1.000 |
| 3 | 0.409 | 0.458 | 0.338 | 0.305 | 0.332 | 0.384 | 0.000 |
| 4 | 0.295 | 0.271 | 0.246 | 0.327 | 0.323 | 0.272 | 0.184 |
| 5 | 0.148 | 0.153 | 0.157 | 0.183 | 0.224 | 0.000 | 0.107 |
| 6 | 0.111 | NA | NA | NA | NA | NA | NA |
| 7 | 0.250 | NA | NA | NA | NA | NA | NA |
| 8 | 0.000 | NA | NA | NA | NA | NA | NA |

Chinese

| $m$ | $n=2$ | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| 4 | 1.000 | 1.000 | 0.746 | 0.712 | 0.75 | 0.689 | 1.00 |
| $4^2$ | 0.834 | 0.867 | 0.862 | 0.869 | 0.88 | 0.907 | 0.87 |
| $4^3$ | 0.822 | 1.000 | NA | NA | NA | NA | NA |
| $4^4$ | 0.833 | NA | NA | NA | NA | NA | NA |

Table 2: Moment estimates $\widehat{\rho}_{nm}$.

| $n$ | $k$ | $m=3$ | | 4 | | 5 | |
|---|---|---|---|---|---|---|---|
| 3 | 1 | 37.42 | 41 | 112.04 | 111 | 192.28 | 193 |
| | 2 | 60.12 | 53 | 73.33 | 81 | 53.31 | 52 |
| | 3 | 8.47 | 12 | 6.63 | 0 | 2.41 | 3 |
| 4 | 1 | 18.99 | 18 | 32.70 | 33 | 57.22 | 57 |
| | 2 | 23.29 | 26 | 21.57 | 21 | 21.92 | 24 |
| | 3 | 3.72 | 2 | 2.68 | 3 | 1.83 | 0 |
| | 4 | 0.00 | 0 | 0.05 | 0 | 0.03 | 0 |
| 5 | 1 | 7.55 | 8 | 11.31 | 12 | 22.31 | 22 |
| | 2 | 8.87 | 8 | 17.09 | 20 | 11.24 | 12 |
| | 3 | 1.58 | 2 | 5.29 | 1 | 1.39 | 1 |
| | 4 | 0.00 | 0 | 0.31 | 1 | 0.05 | 0 |
| | 5 | 0.00 | 0 | 0.00 | 0 | 0.00 | 0 |
| 6 | 1 | 5.02 | 5 | 3.48 | 3 | 5.29 | 5 |
| | 2 | 7.96 | 8 | 4.88 | 6 | 4.60 | 6 |
| | 3 | 2.02 | 2 | 1.53 | 1 | 1.04 | 0 |
| | 4 | 0.00 | 0 | 0.10 | 0 | 0.07 | 0 |
| | 5 | 0.00 | 0 | 0.00 | 0 | 0.00 | 0 |
| | 6 | 0.00 | 0 | 0.00 | 0 | 0.00 | 0 |
| 7 | 1 | 2.95 | 4 | 2.88 | 3 | 2.72 | 3 |
| | 2 | 7.20 | 5 | 3.21 | 3 | 0.27 | 0 |
| | 3 | 2.85 | 4 | 0.85 | 1 | 0.01 | 0 |
| | 4 | 0.00 | 0 | 0.05 | 0 | 0.00 | 0 |
| | 5 | 0.00 | 0 | 0.00 | 0 | 0.00 | 0 |
| | 6 | 0.00 | 0 | 0.00 | 0 | 0.00 | 0 |

Japanese, $\widehat{\gamma}_{nm}$

| $m$ | $n=3$ | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|
| 3 | 0.731 | 0.360 | 0.283 | 0.331 | 0.445 |
| 4 | 0.157 | 0.123 | 0.242 | 0.199 | 0.145 |
| 5 | 0.047 | 0.053 | 0.060 | 0.094 | 0.010 |

Japanese, $\widehat{\rho}_{nm}$

| $m$ | $n=3$ | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|
| 3 | 0.458 | 0.346 | 0.306 | 0.332 | 0.381 |
| 4 | 0.289 | 0.247 | 0.369 | 0.332 | 0.276 |
| 5 | 0.153 | 0.167 | 0.186 | 0.255 | 0.038 |

Table 3: Number of Japanese homonym groups $y_{nmk}$, fitted and observed, (a part of Table 6).

If the homonyms are formed by a common Pitman clustering process, the parameter $\gamma_{nm}$ of $y_{nmk}$ is independent of $n$, and the common values $\gamma_{\cdot m}$ are determined from a common probability $\rho$ by (7). Hence, like the two-way contingency tables, the following three hypotheses are possible.

$$H_1 : \ \gamma_{nm} = \gamma_{\cdot m}, \quad H_2 : \ \gamma_{nm} = \gamma_{n\cdot}, \quad H_3 : \ \gamma_{nm} = \gamma_{\cdot\cdot}$$

Moreover uniformity of $\rho$ is a difenrent alternative for $H_2$ and $H_3$. Among them only the case $H_1$ fit well as shown in **Table 4**. In the other cases the alternatives are strongly significant.

English

| $m$ | $\widehat{\gamma}$ | $\widehat{\rho}$ | d.f. | $p$-value |
|---|---|---|---|---|
| 2 | 0.300 | 0.188 | 3 | 0 |
| 3 | 0.0084 | 0.016 | 2 | .987 |
| all | 0.100 | | 7 | 0 |
| | | 0.025 | | .995 |

Japanese

| $m$ | $\widehat{\gamma}$ | $\widehat{\rho}$ | d.f. | $p$-value |
|---|---|---|---|---|
| 2 | 1.280 | 0.360 | 6 | .784 |
| 3 | 0.503 | 0.401 | 12 | .391 |
| 4 | 0.100 | 0.229 | 17 | .000 |
| 5 | 0.050 | 0.160 | 21 | .997 |
| all | 0.143 | | 62 | 0 |
| | | 0.270 | | 0 |

Chinese

| $m$ | $\widehat{\gamma}$ | $\widehat{\rho}$ | d.f. | $p$-value |
|---|---|---|---|---|
| 4 | 11.912 | 0.735 | 16 | 0.262 |
| $4^2$ | 0.642 | 0.854 | 27 | 0.013 |
| $4^3$ | 0.700 | 0.963 | 2 | 0.000 |

Table 4: Minimum chi-square estimate $\hat{\gamma}$, assuming uniformity within the same $m$.

# Appendix

## Generalized Stirling numbers

Hsu and Shiue (1998) succeeded to unify various extensions of Stirling numbers proposed so far. Their definition was modified slightly by Remmel and Wachs (2004), and the modified one is adopted in this paper. Let Pochhammer's notation of the descending factorial be generalized such that

$$(z|\alpha)_n := \prod_{j=0}^{n-1}(z - j\alpha), \quad (z)_n = (z|1)_n, \quad z^n = (z|0)_n, \quad z^{\overline{n}} = (z|-1)_n.$$

The *generalized Stirling numbers* of the *first and the second kind*, $S^1$ and $S^2$ respectively, are defined by

$$(t - r|\alpha)_n \equiv \sum_{k=0}^{n} S^1_{n,k}(\alpha, \beta, r)\ (t|\beta)_k, \quad (t|\beta)_n \equiv \sum_{k=0}^{n} S^2_{n,k}(\alpha, \beta, r)\ (t - r|\alpha)_k,$$

for any real numbers $\alpha$, $\beta$, $r$. They are orthogonal in the sense that

$$\sum_{k=m}^{n} S^2_{n,k}(\alpha, \beta, r)\ S^1_{k,m}(\alpha, \beta, r) = I[m = n], \quad 1 \le m \le n,$$

and satisfy the recurrence formulas

$$\begin{aligned}
S^1_{n+1,k}(\alpha, \beta, r) =& (k\beta - n\alpha - r)\ S^1_{n,k}(\alpha, \beta, r) + S^1_{n,k-1}(\alpha, \beta, r), \quad 0 \le k \le n+1,\\
S^1_{0,0}(\alpha, \beta, r) =& 1; \quad S^1_{n,k}(\alpha, \beta, r) = 0, \quad \text{if} \quad k < 0 \quad \text{or} \quad k > n,\\
S^2_{n+1,k}(\alpha, \beta, r) =& (k\alpha - n\beta + r)\ S^2_{n,k}(\alpha, \beta, r) + S^2_{n,k-1}(\alpha, \beta, r), \quad 0 \le k \le n+1,\\
S^2_{0,0}(\alpha, \beta, r) =& 1; \quad S^2_{n,k}(\alpha, \beta, r) = 0, \quad \text{if} \quad k < 0 \quad \text{or} \quad k > n.
\end{aligned}$$

In terms of the classical Stirling numbers,

$$\begin{aligned}
S^1_{n,k}(\alpha, \beta, r) &= \sum_{0 \le k \le \ell \le m \le n} \left[ \begin{array}{c} n \\ m \end{array} \right] \left( \begin{array}{c} m \\ \ell \end{array} \right) \left\{ \begin{array}{c} \ell \\ k \end{array} \right\} (-\alpha)^{n-m} (-r)^{m-\ell} \beta^{\ell-k},\\
S^2_{n,k}(\alpha, \beta, r) &= \sum_{0 \le k \le \ell \le m \le n} \left[ \begin{array}{c} n \\ m \end{array} \right] \left( \begin{array}{c} m \\ \ell \end{array} \right) \left\{ \begin{array}{c} \ell \\ k \end{array} \right\} (-\beta)^{n-m} r^{m-\ell} \alpha^{\ell-k}.
\end{aligned}$$

Note that, if $r = 0$ one should read $m = \ell$. These expressions show the relationship between $S^1$ and $S^2$.

The definition induces a family of pmf's $p^1$ and $p^2$ on $\{0, 1, \ldots, n\}$ if $r \neq 0$ and on

$\{1, \ldots, n\}$ if $r = 0$, provided that emerging factors are non-negative:

$$\begin{aligned} p^1(k, n+1;\ t, \alpha, \beta, r) =& \frac{k\beta - r - n\alpha}{t - r - n\alpha} p^1(k;\ n, t;\ \alpha, \beta, r) \\ &+ \frac{t - (k-1)\beta}{t - r - n\alpha} p^1(k-1;\ n, t;\ \alpha, \beta, r), \quad 1 \le k \le n, \\ p^2(k, n+1;\ t, \alpha, \beta, r) =& \frac{k\alpha + r - n\beta}{t - n\beta} p^2(k;\ n, t;\ \alpha, \beta, r) \\ &+ \frac{t - r - (k-1)\alpha}{t - n\beta} p^2(k-1;\ n, t;\ \alpha, \beta, r), \quad 1 \le k \le n. \end{aligned}$$

They will be called *the family of generalized Stirling distributions.*

| $\alpha$ | $\beta$ | $r$ | |
|---|---|---|---|
| $-1$ | 0 | 0 | classical 1st kind (A3) |
| 0 | 1 | 0 | classical 2nd kind (A2) |
| 0 | 0 | $\pm 1$ | Binomial distribution (B) |
| $-1$ | $\gamma$ | 0 | ($\gamma > -1$) Pitman's random partition (A1) |
| $-1$ | 1 | 0 | Lah distribution (A1) |
| $-1$ | 0 | $r$ | $r < 0$, Nishimura and Sibuya 1st kind (B) |
| 0 | 1 | $r$ | $r < 0$, Nishimura and Sibuya 2nd kind (B) |

Table 5: Possible range of parameter values for $p^1(n, k; \alpha, \beta, r)$.

The family defines random walks on the square lattice. A particle starts from $(0, 0)$, if $r \neq 0$, or from $(1, 1)$, if $r = 0$, and moves from $(n, k)$ to

$$\begin{cases} (n+1, k) & \text{with the probability } \dfrac{k\beta - r - n\alpha}{t - r - n\alpha}, \\ (n+1, k+1) & \text{with the probability } \dfrac{t - k\beta}{t - r - n\alpha}, \end{cases}$$

in the case of $p^1$. A similar random walk is defined by $p^2$. $p^1$ (or $p^2$) is the probability distribution of the particle arriving at $(n, k)$ at the $n$-th step.

**Proposition** The $p^1$ is a valid pmf in the following cases.

A $r = 0$,

A1 $t < \beta < \alpha$ & $t \le n\beta$ & $\beta \le n\alpha$; ($\alpha < \beta < t$ & $n\beta \le t$ & $n\alpha \le \beta$;)

A2 $\alpha = 0$ & $0 < n\beta < t$; ($\alpha = 0$ & $t < n\beta < 0$;)

A3 $\beta = 0$ & $\alpha < 0 < t$; ($\beta = 0$ & $t < 0 < \alpha$.)

B $r \neq 0$,

$t < i\beta < r + j\alpha,\ 0 \le i \le j \le n; \quad (r + j\alpha < i\beta < t,\ 0 \le i \le j \le n.)$

Similar result is obtained for $p^2$. The following list shows the relationship of the new families of generalized Stirling distributions with known distributions. The unspecified parameters, restricted to some domain, become the distribution parameters.

# References


[1] Charalambides, C. A. (2005). *Combinatorial Methods in Discrete Distributions*, Wiley, New York, N.Y.

[2] Hoshino, N. (2001). Applying Pitman's sampling formula to microdata disclosure risk assessment, *Journal of Official Statistics*, **17**, 4, 499-520.

[3] Hsu, L. C. and Shiue, P. J-S. (1998). A unidied approach to generalized Stirling numbers, *Advances in Applied Mathematics*, **20**, 366–384.

[4] Johnson, N. L., Kotz, S. and Balakrishnan, N. (1997). *Discrete Multivariate Distributions*, Wiley.

[5] Nishimura, K. and Sibuya, M. (1997). Extended Stirling family of discrete probability distributions, *Communication Statistics, Theory and Method*, **26**, 7 , 1727–1744.

[6] Pitman, J. (1999). Brownian motion, bridge, excursion and meander characterized by sampling at independent uniform times, *Electronic Journal of Probability*, **4**, 11, 1–33.

[7] Pitman, J. (2002). Combinatorial Stochastic Processes, Lecture Notes for St. Flour Course, 207p.

[8] Remmel, J. B. and Wachs, M. L. (2004). Rook theory, generalized Stirling numbers and $(p, q)$-analogues, *The Electronic Journal of Combinatorics*, **11**, 1–48.

[9] Sibata, T. and Shibata, R. (1990) Is word-accent significant in differentiating homonyms in Japanese, English and Chinese? *Mathematical Linguistics*, **17**, 1–11, (in Japanese).

[10] Sibuya, M. (1991) A cluster-number distribution and its application to the analysis of homonyms, *Japan. J. Appl. Statist.*, **20**, 139–153, (in Japanese).

[11] Sibuya, M. and Yamato, H. (2000) Pitman's model of random partitions, *RIMS Kokyuroku (Kyoto University)*, **1240**, 64–73.

[12] Yamato, H., Sibuya, M. and Nomachi, T. (2001). Ordered sample from two-parameter GEM distribution, *Statistics and Probability Letters*, **55**, 1, 19–27.

| $n$ | $k$ | $m=2$ | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 2 | 1 | 7 | 156 | 592 | 875 | 8 | 3 | 2 |
| | 2 | 6 | 109 | 248 | 152 | 1 | 1 | 0 |
| 3 | 1 | 6 | 41 | 111 | 193 | 0 | 0 | 0 |
| | 2 | 3 | 53 | 81 | 52 | 0 | 0 | 0 |
| | 3 | – | 12 | 0 | 3 | 0 | 0 | 0 |
| 4 | 1 | 5 | 18 | 33 | 57 | 0 | 0 | 0 |
| | 2 | 8 | 26 | 21 | 24 | 0 | 0 | 0 |
| | 3 | – | 2 | 3 | 0 | 0 | 0 | 0 |
| | 4 | – | – | 0 | 0 | 0 | 0 | 0 |
| 5 | 1 | 0 | 8 | 12 | 22 | 0 | 0 | 0 |
| | 2 | 1 | 8 | 20 | 12 | 0 | 0 | 0 |
| | 3 | – | 2 | 1 | 1 | 0 | 0 | 0 |
| | 4 | – | – | 1 | 0 | 0 | 0 | 0 |
| | 5 | – | – | – | 0 | 0 | 0 | 0 |
| 6 | 1 | 0 | 5 | 3 | 5 | 0 | 0 | 0 |
| | 2 | 2 | 8 | 6 | 6 | 0 | 0 | 0 |
| | 3 | – | 2 | 1 | 0 | 0 | 0 | 0 |
| | 4 | – | – | 0 | 0 | 0 | 0 | 0 |
| | 5 | – | – | – | 0 | 0 | 0 | 0 |
| 7 | 1 | 0 | 4 | 3 | 3 | 0 | 0 | 0 |
| | 2 | 1 | 5 | 3 | 0 | 0 | 0 | 0 |
| | 3 | – | 4 | 1 | 0 | 0 | 0 | 0 |
| | 4 | – | – | 0 | 0 | 0 | 0 | 0 |
| | 5 | – | – | – | 0 | 0 | 0 | 0 |
| 8 | 1 | 0 | 1 | 1 | 5 | 0 | 0 | 0 |
| | 2 | 2 | 0 | 1 | 2 | 0 | 0 | 0 |
| | 3 | – | 0 | 0 | 0 | 0 | 0 | 0 |
| | 4 | – | – | 0 | 0 | 0 | 0 | 0 |
| | 5 | – | – | – | 0 | 0 | 0 | 0 |

Table 6: Number of Japanese homonym groups with $n$: size, $m$: kinds of accents, $k$: number of clusters. Each row shows $y_{nmk}$, $1 \le k \le \ell,\ \ell = \min(n, m)$. The short hyphens denote impossible entries. For $n \ge 6$, zeros for $k \ge 6$ are unlisted.

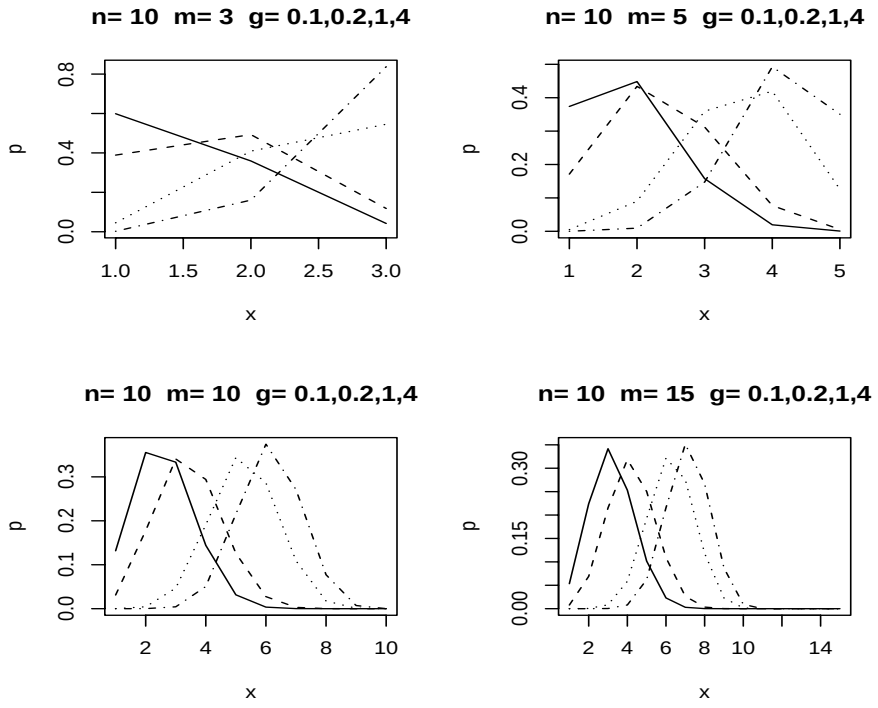


Figure 1: The probability function of $K$ when $m$ and $\gamma$ vary ($n = 10$).

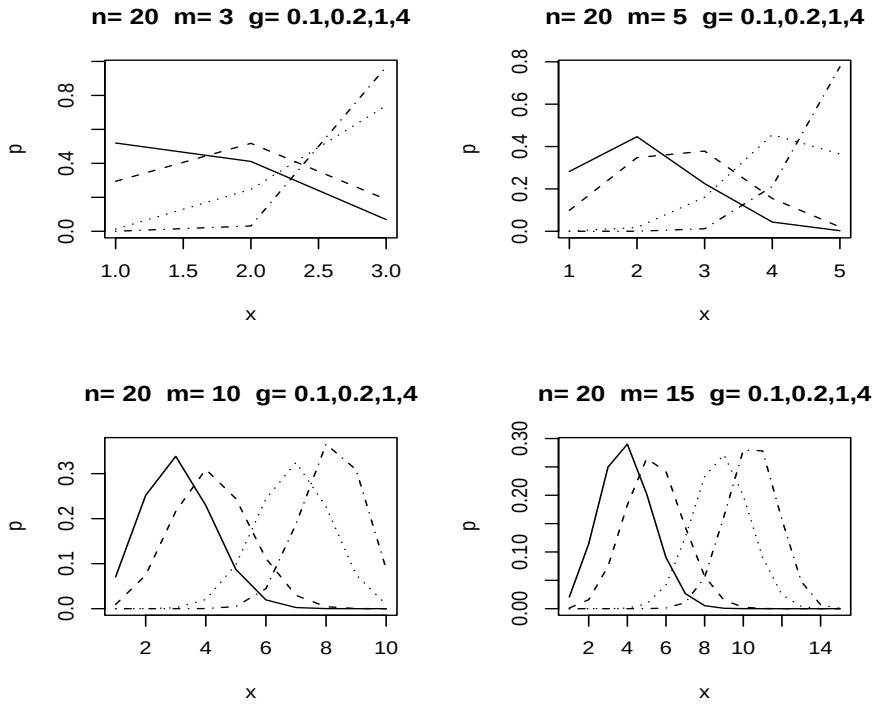


Figure 2: The probability function of $K$ when $m$ and $\gamma$ vary ($n = 20$).

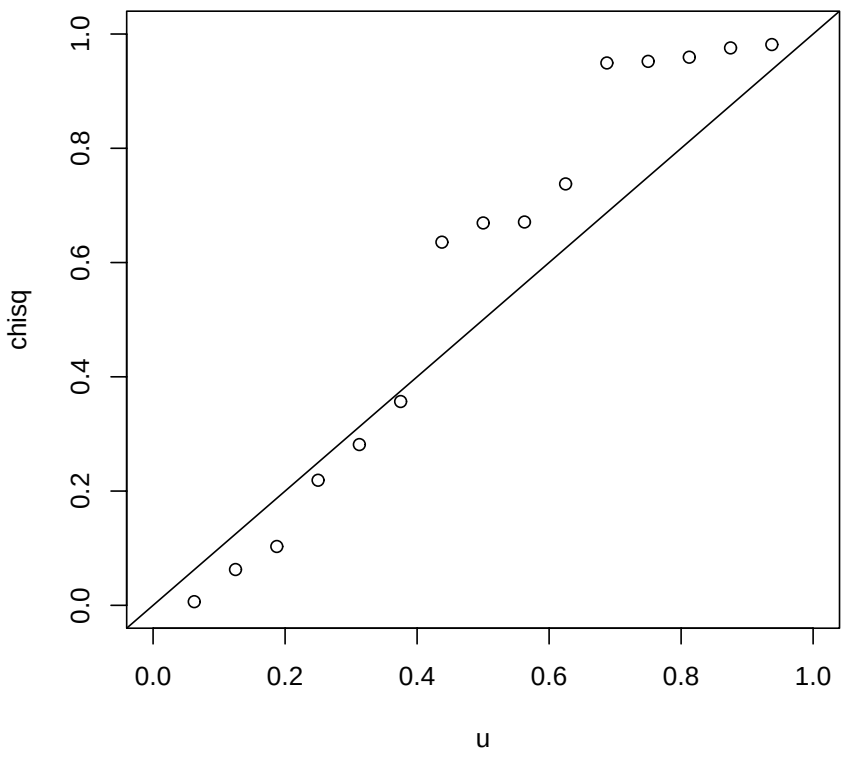


Figure 3: PP-plot of chi-square statistics of Table 3.