GENERALIZED SHIFT TEST METHOD AND METROPOLIS WALK

Sakata, Toshio
Department of Computer Science, Kumamoto University

Nomakuchi, Kentaro
Department of Mathematics, Kochi University

Hayashi, Tadashi
Department of Mathematics, Kumamoto University

https://doi.org/10.5109/13458

出版情報：Bulletin of informatics and cybernetics. 29 (1), pp.1-13, 1997-03. Research Association of Statistical Sciences
バージョン：
権利関係：

Bulletin of Informatics and Cybernetics, Vol. 29, No. 1, 1997

GENERALIZED SHIFT TEST METHOD AND METROPOLIS WALK

By

Toshio SAKATA *, **Kentaro NOMAKUCHI** †and **Tadashi HAYASHI** ‡

Abstract

The concordance number in the generalized shift test method is interpreted as the trace of contingency tables with fixed margins. For the distribution of the concordance number the closeness of the normal approximation and its improvement by the Gram-Charlier approximation are discussed. The Metropolis walk on the set of contingency tables is constructed. The two approximations are compared with the simulation by the Metropolis walk, and then the Gram-Charlier approximation is shown to be extremely close to the distribution of the concordance number.

Key words and phrases: Lexicostatistics, generalized shift test method, contingency tables, Metropolis walk, normal approximation, Gram-Charlier approximation.

1. Introduction

In lexicostatistics one is concerned whether two languages are near or not, in other words, whether they have a common root language or not. One typical idea to confirm it is as follows. A sample of pairs of words with a same meaning is taken from two languages. For example Yasumoto, B. and Honda, M.(1978) selected 100 or 200 pairs of words to verify the nearness between Tokyo dialect and Peking dialect. They, for example, compared "ookii" in Tokyo dialect and "ta" in Peking dialect both of which mean "big," and "kawaita" in Tokyo dialect and "kan" in Peking dialect both of which mean "dry." The number of pairs with the same pronunciation of some part of them, for example the first position, is counted. If this count, i.e., the number of concordant pairs, exceeds the expected number by incidence, it is thought that there is a possibility of the existence of a common root language at some past time and that they are "near." Oswalt, R.L.(1970) introduced the shift test method(STM) to judge whether the concordance number is more than expected or not. Yoshida, T.(1984) generalized the method to supplement some defaults. In the following we give a brief review of Yoshida's method.

* Department of Computer Science, Kumamoto University, Kurokami Kumamoto 860, Japan
E-mail: sakata@toko.ge.kumamoto-u.ac.jp

† Department of Mathematics, Kochi University, Kochi 780, Japan

‡ Department of Mathematics, Kumamoto University, Kumamoto 860, Japan

Suppose that we want to compare a language L_A with a language L_B and that we have a list of N fundamental words $\Gamma_A = \{A_1, \ldots, A_N\}$ and $\Gamma_B = \{B_1, \ldots, B_N\}$ from each language. When a consonant appears first in a word, we call it the first consonant of the word. Let $W = \{w_1, \ldots, w_d\}$ be the collection of all the first consonants which appear in the fundamental words of $\Gamma = \Gamma_A \cup \Gamma_B$. We define a function ϕ as a function which picks up the first consonant of a word. For example, if A="kawaita" of Tokyo dialect which means "dry" then ϕ(A)="k"and if B="kan" of Peking dialect which also means "dry" then ϕ(B)="k," and they have the same first consonant "k." Moreover, if A="nemuru" of Tokyo dialect which means "sleep" then ϕ(A)="n" and if B="inu" of Old Japanese which also means "sleep" then ϕ(B)="n," and they also have the same first consonant. Note that in the last example the first position of "inu" is "i" and it is not a consonant. In this case we are concerned with the pronunciation of "n" which is the consonant to appear first. We define

$$a_i = \sharp\{j : \phi(A_j) = w_i\}, \quad i = 1, \ldots, d,$$

that is, a_i denotes the number of words in Γ_A whose first consonant is w_i. Clearly we have

$$\sum_{i=1}^{d} a_i = N.$$

Similarly we define

$$b_i = \sharp\{j : \phi(B_j) = w_i\}, \quad i = 1, \ldots, d,$$

and then we also have

$$\sum_{i=1}^{d} b_i = N.$$

Let S_N be the symmetric group of degree N. Using a function

$$H(w_1, w_2) = \begin{cases} 1 & \text{if} \quad w_1 = w_2 \\ 0 & \text{otherwise,} \end{cases}$$

we define

$$X = \sum_{i=1}^{N} H(\phi(A_i), \phi(B_{\sigma(i)})),$$

where $\sigma \in S_N$. That is, X is the concordance number among the pairs $(A_1, B_{\sigma(1)})$, ..., $(A_N, B_{\sigma(N)})$. Let x_0 denote the concordance number corresponding with the identity permutation, that is, the one for the original pairing where two words with a same meaning are paired. We then calculate a p-value, i.e., the probability that X becomes greater than or equal to x_0, when σ is uniformly distributed over S_N. If the probability is too small (for example, smaller than 0.01), we think that this suggests the existence of a common root language in the past. Such is an outline of the Yoshida's method, and we call it the generalized shift test method(GSTM). Note that the GSTM improves the STM in the point that the p-value by the GSTM becomes independent of initial arrangements of words.

Yoshida, T.(1984) obtained moments of X and proposed several approximations of the p-values, in which the approximation by the binomial distribution, the normal approximation and the Gram-Charlier approximation are included. However only the approximation by the binomial distribution and the normal approximation are treated there for real data. We discuss the Gram-Charlier approximation in more details. He also suggested that the normal approximation for the GSTM may be good because the skewness and the kurtosis are very near to those of the normal. But he argued nothing about the closeness of these approximations.

The main purpose of this paper is to discuss the closeness of the normal approximation and its improvement by the Gram-Chariler approximation. The notion of the concordance number of the GSTM can be restated in terms of contingency tables with fixed margins, and then the distribution of X becomes theoretically tractable. In this framework, we show the following; (1) The exact distribution is easily simulated by a Metropolis walk on the set of contingency tables; (2) The normal approximation is sufficiently good for testing statistically with usual significance levels, but it gives too small estimates for the small upper probabilities; (3) The Gram-Charlier approximation improves the normal approximation and fits extremely well even for the small upper probabilities.

2. Concordance number and contingency table

Let Ω be the set of all $d \times d$ contingency tables with the fixed row sums $\{a_1, \ldots, a_d\}$ and the fixed column sums $\{b_1, \ldots, b_d\}$. For $\theta \in \Omega$, let θ_{ij} be the (i,j) cell's count. The following proposition summarizes the relation of the two notions of concordance number and contingency table, the proof of which is easily obtained by a combinatorial argument.

PROPOSITION 2.1. *The concordance number can be interpreted as the diagonal sum (that is, the trace) of θ in Ω with the specified row sums and column sums. Furthermore $\theta \in \Omega$ is distributed as a generalized hypergeometric distribution, that is,*

$$P(\theta) = \frac{\prod_{i=1}^{d} a_i! \prod_{j=1}^{d} b_j!}{N! \prod_{i=1}^{d} \prod_{j=1}^{d} \theta_{ij}!}.$$

Table 1 illustrates the proposition. The second column of the table shows that among the b_1 words of Γ_B whose first consonant is w_1, θ_{11}, θ_{21} and θ_{31} words correspond to the words of Γ_A whose first consonant is w_1, w_2 and w_3 respectively. Other columns are interpreted in the same way. This table corresponds to the Γ_A and Γ_B.

Table 1. 3×3 contingency table for Γ_A and Γ_B.

$\phi(\Gamma_A)\backslash\phi(\Gamma_B)$	w_1	w_2	w_3	total
w_1	θ_{11}	θ_{12}	θ_{13}	a_1
w_2	θ_{21}	θ_{22}	θ_{23}	a_2
w_3	θ_{31}	θ_{32}	θ_{33}	a_3
total	b_1	b_2	b_3	N

In the similar way, we have another table $\theta(\sigma)$ for $\sigma \in \Omega$ from Γ_A and $\Gamma_B^\sigma = \{B_{\sigma(1)}, \ldots, B_{\sigma(3)}\}$. When σ is uniformly distributed in S_3, we have the generalized hypergeometric distribution of $\theta = \theta(\sigma)$ such as

$$P(\theta) = \frac{\prod_{i=1}^3 a_i! \prod_{j=1}^3 b_j!}{N! \prod_{i=1}^3 \prod_{j=1}^3 \theta_{ij}!}.$$

The concordance number is given by $X = \sum_{i=1}^3 \theta_{ii}$.

Next we consider the distribution of X. Our primary concern about the distribution of $X = \sum_{i=1}^d \theta_{ii}$ is in the accuracy of the normal approximation to the distribution. In reality the number of fundamental words used in comparisons are N=100 or 200 (see for example, Yasumoto, B. and Honda, M.(1978)). Moreover, if we regard that some groups of consonants, for example, "h" and "p" or "b" and "v", have the same pronunciation, the number of all consonants is d=21. If this is the case, a_i and b_i are about N/d =5 or 10 and hence contingency tables tend very sparse. Thus, as is well known, the normal approximation of a specific cell count θ_{ij} might not be good. But we are concerned in the distribution of the diagonal sum $X = \sum_{i=1}^d \theta_{ii}$ which might have a good approximation to the normal. We will give the moment formulas up to the fourth order of X, and will compare the kurtosis and skewness of the distribution with those of a normal distribution.

The following lemma is well known, but we give a proof since it is instructive to get the moments of any order.

LEMMA 2.1.

$$E\{\theta_{ij}\} = \frac{a_i b_j}{N}, \quad 1 \le i, j \le d.$$

PROOF.

$$\begin{aligned}
E\{\theta_{ij}\} &= \sum_{\theta \in \Omega} \theta_{ij} P(\theta) \\
&= \sum_{\theta \in \Omega} \theta_{ij} \frac{\prod_{s=1}^d a_s! \prod_{t=1}^d b_t!}{N! \prod_{s=1}^d \prod_{t=1}^d \theta_{st}!} \\
&= \sum_{\theta \in \Omega'} \frac{a_i! b_j!}{\theta_{i1}! \cdots (\theta_{ij}-1)! \cdots \theta_{id}!} \frac{\prod_{s \neq i} a_s! \prod_{t \neq j} b_t!}{N! \prod_{s \neq i} \prod_{t=1}^d \theta_{st}!} \\
&= \frac{a_i b_j}{N} \sum_{\theta \in \Omega'} \frac{(a_i-1)!(b_j-1)!}{\theta_{i1}! \cdots (\theta_{ij}-1)! \cdots \theta_{id}!} \frac{\prod_{s \neq i} a_s! \prod_{t \neq j} b_t!}{(N-1)! \prod_{s \neq i} \prod_{t=1}^d \theta_{st}!} \\
&= \frac{a_i b_j}{N},
\end{aligned}$$

where $\Omega' = \{\theta \in \Omega : \theta_{ij} \ge 1\}$.

We obtain the following moments formulas in a similar way as Lemma 2.1.

PROPOSITION 2.2.

$$E\{X\} = \sum_{i=1}^d \frac{a_i b_i}{N}.$$

PROPOSITION 2.3. *For different integers i, j, k, l, we have the following moment formulas.*

$$
\begin{aligned}
E\{\theta_{ii}\theta_{jj}\} &= \frac{a_i b_i a_j b_j}{N(N-1)}, \\
E\{\theta_{ii}^2\} &= \frac{a_i(a_i-1)b_i(b_i-1)}{N(N-1)} + E\{\theta_{ii}\}, \\
E\{\theta_{ii}\theta_{jj}\theta_{kk}\} &= \frac{a_i b_i a_j b_j a_k b_k}{N(N-1)(N-2)}, \\
E\{\theta_{ii}^2\theta_{jj}\} &= \frac{a_i(a_i-1)b_i(b_i-1)a_j b_j}{N(N-1)(N-2)} + E\{\theta_{ii}\theta_{jj}\}, \\
E\{\theta_{ii}^3\} &= \frac{a_i(a_i-1)(a_i-2)b_i(b_i-1)(b_i-2)}{N(N-1)(N-2)} + 3E\{\theta_{ii}^2\} - 2E\{\theta_{ii}\}, \\
E\{\theta_{ii}\theta_{jj}\theta_{kk}\theta_{ll}\} &= \frac{a_i b_i a_j b_j a_k b_k a_l b_l}{N(N-1)(N-2)(N-3)}, \\
E\{\theta_{ii}^2\theta_{jj}\theta_{kk}\} &= \frac{a_i(a_i-1)b_i(b_i-1)a_j b_j a_k b_k}{N(N-1)(N-2)(N-3)} + E\{\theta_{ii}\theta_{jj}\theta_{kk}\}, \\
E\{\theta_{ii}^3\theta_{jj}\} &= \frac{a_i(a_i-1)(a_i-2)b_i(b_i-1)(b_i-2)a_j b_j}{N(N-1)(N-2)(N-3)} \\
&\quad + 3E\{\theta_{ii}^2\theta_{jj}\} - 2E\{\theta_{ii}\theta_{jj}\}, \\
E\{\theta_{ii}^2\theta_{jj}^2\} &= \frac{a_i(a_i-1)b_i(b_i-1)a_j(a_j-1)b_j(b_j-1)}{N(N-1)(N-2)(N-3)} \\
&\quad + E\{\theta_{ii}^2\theta_{jj}\} + E\{\theta_{ii}\theta_{jj}^2\} - E\{\theta_{ii}\theta_{jj}\}, \\
E\{\theta_{ii}^4\} &= \frac{a_i(a_i-1)(a_i-2)(a_i-3)b_i(b_i-1)(b_i-2)(b_i-3)}{N(N-1)(N-2)(N-3)} \\
&\quad + 6E\{\theta_{ii}^3\} - 11E\{\theta_{ii}^2\} + 6E\{\theta_{ii}\}, \\
E\{X^2\} &= \sum_{i=1}^{d} E\{\theta_{ii}^2\} + \sum_{i\neq j} E\{\theta_{ii}\theta_{jj}\}, \\
E\{X^3\} &= \sum_{i=1}^{d} E\{\theta_{ii}^3\} + 3\sum_{i\neq j} E\{\theta_{ii}^2\theta_{jj}\} + \sum\nolimits_1 E\{\theta_{ii}\theta_{jj}\theta_{kk}\}, \text{ and} \\
E\{X^4\} &= \sum_{i=1}^{d} E\{\theta_{ii}^4\} + 4\sum_{i\neq j} E\{\theta_{ii}^3\theta_{jj}\} + 3\sum_{i\neq j} E\{\theta_{ii}^2\theta_{jj}^2\} \\
&\quad + 6\sum\nolimits_1 E\{\theta_{ii}^2\theta_{jj}\theta_{kk}\} + \sum\nolimits_2 E\{\theta_{ii}\theta_{jj}\theta_{kk}\theta_{ll}\},
\end{aligned}
$$

where none of i, j, k are equal in $\sum_1$ and none of i, j, k, l are equal in $\sum_2$.

To measure the closeness between the distribution of X and the normal, we can calculate the skewness $s = E\{(X-\mu)^3\}/\sigma^3$ and the kurtosis $k = E\{(X-\mu)^4\}/\sigma^4$ by using these moment formulas, where μ is the mean and σ^2 is the variance. For the special cases of $N = d^2$ and the row sums $\{d, d, \ldots, d\}$ and the column sums $\{d, d, \ldots, d\}$, these quantities are obtained as functions of d,

$$s = \frac{(d-2)\sqrt{d+1}}{d^2-2},$$
$$k = \frac{(d+1)(3d^4-2d^3-18d^2+19d+6)}{d(d^2-2)(d^2-3)}.$$

These become asymptotically 0 and 3, respectively, which suggests the fitness of the normal approximation. As we will show later in Section 4 and 5, the Gram-Charlier approximation is much more good. The following Table 2 lists the values of skewnesses and kurtosises for $d = 3, 4, \ldots, 10$.

Table 2. Skewness and kurtosis of the concordance number, $N = d^2$, the row sums $\{d, ..., d\}$, and the column sums $\{d, ..., d\}$.

d	3	4	5	6	7	8	9	10
skewness	0.286	0.319	0.319	0.311	0.301	0.290	0.280	0.271
kurtosis	2.857	2.981	3.026	3.045	3.052	3.055	3.055	3.054

3. Metropolis walk on the set of contingency tables

Let $\Omega = \Omega(\{a_i\}, \{b_i\})$ be the set of all $d \times d$ contingency tables with the fixed row sums $\{a_i\}$ and the fixed column sums $\{b_i\}$. To simulate the generalized hypergeometric distribution on the set Ω, we utilize a Metropolis walk on it which is a Markov chain with the objective distribution in the limit. When the Metropolis walk reaches to the limit distribution, we can generate a random element with the limit distribution. But, tracing the walk and recording it at intervals of some steps, we can expect to get a series of random elements with the objective distribution. Such a walk on the set of contingency tables is argued in Diaconis, P. and Holmes, P.(1995) where the limit distribution is a uniform distribution. For a more general framework of the Metropolis walk see Besag, J., Green, P., Higdon, D. and Mengersen,K.(1995) and for some examples of the calculation of the convergence rate see Diaconis, P. and Hanlon, P.(1992).

In the following we state the construction process of the Metropolis walk.

1. Select an initial starting $d \times d$ contingency table θ.

2. In θ, select randomly two different rows i_1 and i_2 and two different columns j_1 and j_2.

3. Choose a candidate, say θ', of the next state (contingency table) as follows. Throw a fair coin. If a head occurs, add 1 to the cells (i_1, j_1) and (i_2, j_2), and subtract 1 from the cells (i_1, j_2) and (i_2, j_1). If a tail occurs, subtract 1 from the cells (i_1, j_1) and (i_2, j_2), and add 1 to the cells (i_1, j_2) and (i_2, j_1). The resulting contingency table is the candidate θ'.

4. If the θ' is illegal (that is, goes outside of Ω) then the walk keeps staying at the present state. Otherwise the walk moves to the state θ' with an acceptance

probability $\min\{1, \pi(\theta') \,/\, \pi(\theta)\}$ where π denotes the objective generalized hypergeometric probability. Set $\theta = \theta'$ and go to Step 2.

Using this Metropolis walk we are able to simulate various aspects of the distribution of the sum of diagonals, that is, the concordance number. To check the fit of the simulation by the Metropolis walk, here we consider a special case of $a_1 = \cdots = a_d = b_1 = \cdots = b_d = d$, $d = 3, 4, ..., 10$. Table 3 lists the moments of X obtained by the Metropolis walk simulation where μ_3 and μ_4 are the central moments of the third and the fourth order, respectively. Note that in this case $E(X) = d$. In each walk initial 5000 steps were discarded and then 50000 steps were recorded in every 120 steps by which the empirical distribution and moments were calculated. Listed in the lower half of the table are the exact moments obtained by Proposition 2.3. The closeness of the figures in the upper and lower halves of the table indicates that the walks have attained the limit distribution, though the closeness between moments is a week evidence for the closeness between distributions. This encourages us to study the various aspects of X by this method. Note that we will more closely check the convergence of the distribution in the later sections.

Table 3. Moments of X,
upper half : moments by simulation,
lower half : exact moments.

	μ	σ^2	μ_3	μ_4
$d = 3$	3.007	2.251	0.939	14.355
	3.000	2.250	0.964	14.464
$d = 4$	3.992	3.170	1.883	30.106
	4.000	3.200	1.829	30.523
$d = 5$	5.011	4.151	2.714	52.942
	5.000	4.167	2.717	52.536
$d = 6$	6.026	5.158	3.294	79.495
	6.000	5.143	3.630	80.525
$d = 7$	6.999	6.085	4.743	113.41
	7.000	6.125	4.561	114.51
$d = 8$	8.008	7.148	5.585	156.75
	8.000	7.111	5.505	154.48
$d = 9$	9.007	8.074	5.280	193.02
	9.000	8.100	6.459	200.36
$d = 10$	9.950	8.905	6.365	235.82
	10.00	9.091	7.421	252.43

4. Normal approximation

In Section 2 we saw that the skewness and kurtosis of the null distribution of the concordance number were near to those of the normal distribution. In this section, we

discuss how close the upper probabilities of X are to those of the normal distribution. We obtain the upper probabilities through three methods, the Metropolis simulation, the normal approximation, and the Gram-Charlier approximation by using the moments up to the fourth order obtained in Section 2. Note that in this paper both approximations are always adjusted by continuity correction. In this Metropolis simulation the number of replication is 900000. Except this we use the same parameters as in Section 3.

Table 4. Upper probabilities.

d	x	Metro. sim.	Norm. appr.	G.-C. appr.	Exact dist.
$d=3$	5	0.16088	0.15866	0.16154	0.16131
	6	0.04870	0.04779	0.05643	0.04880
	7	0.01652	0.00982	0.01447	0.01667
	8	0.00061	0.00135	0.00256	0.00059
$d=4$	6	0.19601	0.20087	0.19686	0.19645
	7	0.08653	0.08113	0.08893	0.08658
	8	0.03169	0.02520	0.03398	0.03174
	9	0.00954	0.00594	0.01061	0.00962
$d=5$	8	0.11390	0.11034	0.11498	0.11373
	9	0.05053	0.04321	0.05267	0.05043
	10	0.01970	0.01374	0.02112	0.01953
	11	0.00667	0.00353	0.00718	0.00662
$d=6$	9	0.13574	0.13514	0.13678	***
	10	0.06757	0.06137	0.06985	***
	11	0.03060	0.02361	0.03229	***
	12	0.01239	0.00765	0.01326	***
	13	0.00449	0.00208	0.00472	***
$d=8$	11	0.17171	0.17425	0.17157	***
	12	0.09913	0.09468	0.09992	***
	13	0.05249	0.04575	0.05429	***
	14	0.02591	0.01958	0.02735	***
	15	0.01181	0.00739	0.01262	***
	16	0.00496	0.00246	0.00525	***
$d=10$	14	0.12539	0.12286	0.12516	***
	15	0.07362	0.06779	0.07470	***
	16	0.04063	0.03407	0.04210	***
	17	0.02118	0.01555	0.02229	***
	18	0.01043	0.00643	0.01099	***
	19	0.00488	0.00241	0.00500	***

Table 4 lists values x and the upper probabilities $P\{X \geq x\}$ whose values are approximately from 1% to 10%. We have the exact upper probabilities only for $d = 3, 4, 5$ because for $d \geq 6$ the calculation becomes untractable due to the explosion of the number of cases. The normal approximation is given by $1 - \Phi(z)$, $z = (x - \mu - 0.5)/\sigma$, where

μ and σ^2 are the mean and the variance for the exact distribution (Table 3) and Φ is the standard normal distribution function. The Gram-Charlier approximation up to the fourth order(for the definition see Kendall, M.G. and Stuart, A.(1977)) is given by

$$1 - \Phi(z) + \phi(z)\{\frac{1}{6}s(z^2 - 1) + \frac{1}{24}(k - 3)(z^3 - 3z)\},$$

where s and k are the skewness and kurtosis (Table 2), respectively, and ϕ is the standard normal density function. The closeness of the figures in the third column and the sixth column for $d \leq 5$ indicates that the Metropolis walk reaches to the stationary generalized hypergeometric distribution, so the simulation precisely generates the distribution of the concordance number. Table 4 indicates that the normal approximation is a little rough, but the Gram-Charlier approximation is very good, especially for large d.

5. P-values for real data

In this section, we will give the upper probabilities and p-values by the GSTM for the concordance numbers between Old Japanese language and some other languages. The concordance numbers calculated here are between (1) Old Japanese and Korean in the middle age, (2) Old Japanese and Vietnamese, (3) Old Japanese and Indonesian and (4) Old Japanese and Tahitian. We used 200 basic words selected from each languages by Yasumoto, B. and Honda, M.(1978). These pairs of languages are summarized in Table 5, 6, 7, and 8. They treat the consonants in the following groups as having the same pronunciation; (1) h, p, b, f, v and x, (2) t, c, tf, ts, s, d and z, (3) k, g, q and nq, and (4) l and r. We also follow their treatment.

Table 5. Contingency table for Old Japanese and Korean.

	Korean								
	b	c	g	l	m	n	y	*	
b	10	6	6	0	3	7	0	1	33
c	10	27	4	1	5	9	1	0	57
g	9	11	9	3	6	6	0	4	48
l	0	2	0	0	0	0	0	0	2
m	4	6	7	2	3	1	1	1	25
n	3	2	3	1	4	3	0	1	17
w	2	4	1	1	0	2	0	0	10
y	2	1	1	0	1	2	0	1	8
	40	59	31	8	22	30	2	8	200

Table 6. Contingency table for Old Japanese and Vietnamese.

	Vietnamese								
	b	c	g	l	m	n	y	*	
b	5	7	2	8	6	5	0	0	33
c	12	29	4	5	3	3	0	1	57
g	2	28	5	1	3	8	1	0	48
l	0	0	0	1	1	0	0	0	2
m	6	11	1	0	2	5	0	0	25
n	0	8	2	1	1	5	0	0	17
w	1	7	0	0	0	2	0	0	10
y	3	4	0	0	0	1	0	0	8
	29	94	14	16	16	29	1	1	200

Table 7. Contingency table for Old Japanese and Indonesian.

	Indonesian									
	b	c	g	l	m	n	w	y	*	
b	14	5	4	7	3	0	0	0	0	33
c	12	27	8	5	3	1	0	1	0	57
g	13	13	10	4	4	2	1	0	1	48
l	1	1	0	0	0	0	0	0	0	2
m	10	6	4	3	2	0	0	0	0	25
n	2	4	3	3	2	2	0	1	0	17
w	1	3	4	1	0	1	0	0	0	10
y	3	2	1	0	2	0	0	0	0	8
	56	61	34	23	16	6	1	2	1	200

Table 8. Contingency table for Old Japanese and Tahitian.

	Tahitian						
	b	c	l	m	n	*	
b	17	7	3	0	3	3	33
c	17	16	11	9	0	4	57
g	15	11	8	3	4	7	48
l	0	1	1	0	0	0	2
m	5	7	1	1	5	3	25
n	10	1	2	1	1	2	17
w	2	2	0	3	0	3	10
y	3	3	0	2	0	0	8
	69	48	26	23	11	23	200

Yasumoto, B. and Honda, M.(1978) give the p-values by the STM, but we here give the p-values by the GSTM. Table 9 lists the upper probabilities obtained by three methods, the Metropolis simulation, the normal approximation, and the Gram-Charlier approximation.

The concordance numbers obtained from the actual contingency tables (Table 5, 6, 7, 8) are underlined in the second column, the upper probabilities for which are the p-values. In the Metropolis walk, all the effective step numbers are taken 900000 and the effective steps are chosen in every 120 steps. Initial 5000 steps are discarded in each walk.

In table 9, first of all, it should be noted that all the probabilities by the Metropolis simulation is reliable at least up to three places of decimals because the values were stable from 700000 steps to 900000 in our simulations. Table 9 indicates that the p-values by the GSTM for Old Japanese vs. Indonesian and for Old Japanese vs. Korean in the middle age are extremely small. This corresponds to the result in Yasumoto, B. and Honda, M.(1978) that the former pair's p-value by the STM is less than 0.001 and the latter is less than 0.0001. Note that the concordance numbers given by us are slightly different from that given in Yasumoto, B. and Honda, M.(1978). This may be because of some ambiguity in the list of selected words in Yasumoto, B. and Honda, M.(1978). From these differences we could not compare precisely these p-values.

Table 9. Upper probabilities for the concordance numbers.

	x	Metro. sim.	Norm. appr.	G.-C. appr.
Korean	43	0.11742	0.11561	0.11705
	44	0.08475	0.08214	0.08457
	45	0.05964	0.05655	0.05949
	46	0.04078	0.03771	0.04074
	47	0.02713	0.02433	0.02716
	48	0.01745	0.01519	0.01761
	49	0.01096	0.00917	0.01110
	52	0.00234	0.00164	0.00234
Vietnamese	46	0.11471	0.11333	0.11465
	47	0.08093	0.07869	0.08085
	48	0.05521	0.05274	0.05530
	49	0.03631	0.03410	0.03667
	50	0.02344	0.02125	0.02356
	51	0.01447	0.01276	0.01467
	52	0.00878	0.00737	0.00884
Indonesian	44	0.12910	0.12881	0.12976
	45	0.09354	0.09247	0.09440
	46	0.06616	0.06433	0.06684
	47	0.04550	0.04335	0.04604
	48	0.03055	0.02828	0.03085
	49	0.01993	0.01785	0.02010
	50	0.01279	0.01089	0.01272
	55	0.00085	0.00055	0.00083
Tahiti	35	0.11343	0.11298	0.11444
	36	0.07721	0.07542	0.07778
	37	0.05043	0.04823	0.05093
	38	0.03170	0.02952	0.03211
	39	0.01939	0.01727	0.01948
	40	0.01129	0.00966	0.01137

About the closeness between the Metropolis simulation and two approximations, these are considerably close, so both of the two approximations are good. More closely looking at the table, however, we can see that in almost all cases the accuracy of the normal approximation is up to two places of decimals, while that of the Gram-Charlier is up to three places of decimals. For example, in Old Japanese and Vietnamese the p-value by the simulation is 0.08093 and that by the Gram-Charlier approximation is also 0.08085. On the other hand the p-values by the normal approximation is 0.07869. We also see that the Gram-Charlier approximation is much better than the normal approximation for the small upper probability. For example, in old Japanese and Korean

the p-value by the simulation is 0.00234 and that by the Gram-Charlier approximation is also 0.00234. On the other hand the p-values by the normal approximation is 0.00164 and a little far from the one by the Metropolis simulation. It is important that this feature is seen in all the upper probabilities in Table 9. Table 10 lists the very small upper probabilities. We see that for these small probabilities the Gram-Charlier approximation is clearly better than the normal approximation.

Table 10. Approximations for small probabilities .

	x	Metro. sim.	Norm. appr.	G.-C. appr.
Korean	53	0.00137	0.00087	0.00131
	54	0.00076	0.00044	0.00071
	55	0.00040	0.00022	0.00037
Vietnamese	53	0.00517	0.00410	0.00516
	54	0.00293	0.00220	0.00291
	55	0.00163	0.00113	0.00159
Indonesian	55	0.00085	0.00055	0.00083
	56	0.00044	0.00027	0.00044
	57	0.00023	0.00013	0.00022
Tahiti	41	0.00634	0.00516	0.00637
	42	0.00343	0.00263	0.00343
	43	0.00176	0.00128	0.00177

From the definition of the concordance number X, we can expect that its distribution has a slightly heavy tail on right, which is confirmed by the fact that the values by the Metropolis simulation become a little larger than those of the normal approximation in Table 9 and Table 10. In our opinion, using the moments up to the fourth order, the Gram-Charlier approximation reduces the influence of the positive skewness.

In conclusion, the GSTM improves the default of the STM, the dependency to an initial arrangement of fundamental words, and it can be used as a recommendable substitute for the STM. To investigate theoretically the GSTM, it is useful to treat them in the framework of contingency tables. The p-values by the GSTM are obtained through three methods, the Metropolis simulation, the normal approximation and the Gram-Charlier approximation. The Metropolis simulation gives reliable estimates of the upper probabilities of the concordance number. The normal approximation is good for the practical purpose, but it gives too small estimates for extremely small upper probabilities. On the other hand the Gram-Charlier approximation up to the fourth order is much better than the normal approximation and gives much more accurate estimates even for the very small upper probabilities. We recommend the Gram-Charlie approximation.

Acknowledgements

We wish to express our thanks to Professor Tomoyuki Yoshida for introducing us to this topic and to Professor Takashi Yanagawa for his valuable comments and

encouragement.

References

Besag, J., Green, P., Higdon, D. and Mengersen,K. (1995), *Bayesian computation and stochastic systems.* Statistical Science, **10**, No. 1, 3-66.

Diaconis, P. and Hanlon, P. (1992), *Eigen analysis for some examples of the Metropolis algorithm. Contemporary Mathematics*, **138**, 99-117.

Diaconis, P. and Holmes, S. (1995), *Three examples of Monte-Carlo Markov chain: at the interface between statistical computing, computer science and statistical mechanics.* Preprint .

Kendall, M.G. and Stuart, A. (1977), *The Advanced Theory of Statistics.* Vol.1, C. Griffin, London.

Oswalt, R.L. (1970), *The detection of remote linguistic relationships.* Computer studies in the Humanities and Verbal Behavior.

Yasumoto, B. and Honda, M. (1978), *"Nihongo no Tanjyou,"* Daisyukan Shoten, In Japanese.

Yoshida, T. (1984), *"Gengokan no kyori to shift hou,"*, Suri Kagaku (Mathematical Sciences), **258**, 37-42. In Japanese.

Received December 28, 1996
Revised January 20, 1997