SHIFTED BINOMIAL APPROXIMATIONS FOR EWENS SAMPLING FORMULA

Yamato, Hajime
Kagoshima University : Professor emeritus

https://doi.org/10.5109/2232329

出版情報：Bulletin of informatics and cybernetics. 49, pp.81-88, 2017-12. Research Association of Statistical Sciences
バージョン：
権利関係：

SHIFTED BINOMIAL APPROXIMATIONS FOR EWENS SAMPLING FORMULA

by

Hajime Yamato

Reprinted from the Bulletin of Informatics and Cybernetics
Research Association of Statistical Sciences, Vol.49

FUKUOKA, JAPAN
2017

Bulletin of Informatics and Cybernetics, Vol. 49, 2017

SHIFTED BINOMIAL APPROXIMATIONS FOR EWENS SAMPLING FORMULA

By

Hajime Yamato*

Abstract

The Ewens sampling formula is well-known as the distribution of a random partition of the set of integers $\{1, 2, \ldots, n\}$. For the number K_n of distinct components of the formula, K_n has the asymptotic normality and its distribution is approximated by the Poisson distribution (see, for example, Arratia et al. (2003)). But, there is no research on the relation between K_n and the binomial distribution, as far as the author knows. We give the approximations to the distribution of K_n by using the shifted bimonial distribution and several examples by illustration.

Key Words and Phrases: Approximate distribution, Bernoulli random variable, Binomial distribution, Ewens sampling formula, Shifted distribution, Krawtchouk polynomal.

1. Introduction

Ewens (1972) discovered a distribution of a random partition of the set of integers $\{1, 2, \ldots, n\}$, partially intuitively and the distribution is well-known as the Ewens sampling formula. It was derived exactly by Antoniak (1974), using Ferguson's Dirichlet process (Ferguson (1974)). The formula appears in many statistical contexts. For example, Bayesian statistics, pattern of communication and genetics. There are many works on the Ewens sampling formula and the related formula. For the Ewens sampling formula, the number K_n of components has the distribution whose probability function given by $P(K_n = k) =\mid s(n,k) \mid \theta^k/\theta^{[n]} \quad (k = 1, 2, \ldots, n)$, where $\theta > 0$, $\theta^{[n]} = \theta(\theta+1)\cdots(\theta+n-1)$ and $\mid s(n,k) \mid$ is the signless Stirling number of the first kind (see, for example, Johnson et al. (1997) and Arratia et al. (2003)).

It is well-known that K_n has the asymptotic normality (see, for example, Johnson et al. (1997; Chapter 41) and Arratia et al. (2003; Section 5.2)). Since the mean and variance of K_n is written using the digamma function and trigamma function and these function are included in the programming language R, the appropriate approximation to the distribution $\mathcal{L}(K_n)$ of K_n are obtained using R (Yamato et al. (2015)).

The Poisson approximation to the distribution of the number K_n of component are studied by Arratia et al. (2000) in detail with respect to the logarithmic combinatorial structure including Ewens sampling formula. Differently from Arratia et al. (2000), Yamato (2017) approaches to the problem of Poisson approximation to $\mathcal{L}(K_n)$ by using the sum of independent Bernoulli random variables.

Whereas, there is no research on the binomial approximation to $\mathcal{L}(K_n)$. There are researches on the binomial approximations to the distribution of the sum of independent

* Emeritus of Kagoshima University, Take 3-32-1-708, Kagoshima 890–0045, Japan

Bernoulli random variables. Using these results, we give the approximations to $\mathcal{L}(K_n)$ by the shifted binomial distribution.

In Section 2, we quote the previous researches on the binomial approximations to the distribution of the sum of independent but non-identically distributed Bernoulli random variables. Using them, in Section 3 we give the approximations to $\mathcal{L}(K_n)$ by the shifted binomial distribution, which are illustrated by several examples.

2. Sum of independent binomial random variables

Let $X_1, X_2, \ldots, X_n$ be independent Bernoulli random variables with $P(X_j = 1) = p_j$ and $P(X_j = 0) = 1 - p_j$ $(n = 1, 2, \ldots, n)$. We put

$$S_n = \sum_{j=1}^{n} X_j, \quad \lambda_n = \sum_{j=1}^{n} p_j, \quad \lambda_{k,n} = \sum_{j=1}^{n} p_j^k \ (k = 2, 3), \quad \bar{p}_n = \frac{\lambda_n}{n}. \tag{1}$$

2.1. Approximation I

As the approximation to the distribution $\mathcal{L}(S_n)$ of S_n, Ehm (1991) gives the following binomial distribution with parameters n and $\bar{p}_n$:

$$B_N(n, \bar{p}_n). \tag{2}$$

The mean of the approximation $B_N(n, \bar{p}_n)$ is equal to the mean $E(S_n)$.

2.2. Approximation II

Barbour et al. (1992; p.190) and Soon (1996) give the approximations $B_N(n', p')$ whose mean is equal to the mean $E(S_n)$ and variance is approximately equal to the variance $V(S_n)$. Barbour et a. (1992) takes $n' = \lfloor \lambda_n^2/\lambda_{2,n} \rfloor$ and $p' = \lambda_n/n'$, where $\lfloor x \rfloor$ represents the integral part of x. Being a little different from them, Soon (1996) takes $n' = \lfloor \lambda_n^2/\lambda_{2,n} + 0.5 \rfloor$ and $p' = \lambda_n/n'$ in the case where $\lambda_{2,n}/\lambda_n$ is bounded away from 1. While, the model of the next section does not satisfy this condition. Therefore, as Approximation II, following Barbour et al. (1992; p.190) we take

$$B_N(n', p') \quad (n' = \lfloor \lambda_n^2/\lambda_{2,n} \rfloor \ \text{ and } \ p' = \lambda_n/n'). \tag{3}$$

2.3. Approximation III

Čekanavičius et al. (2009) give the approximation whose first three moments are approximately equal to them of S_n, respectively. Let the random variable Y have the shifted distribution given by the Binomial distribution $B_N(n, p)$ with the shift s, that is $Y \sim s + B_N(n, p)$. As the approximation to S_n, the parameters of Y are determined as follows. By approximately equaling the first three moments of S_n and $Y (\sim s + B_N(n, p))$, the following p^*, n^* and s^* are obtained as the solutions of p, n and s, respectively.

$$p^* = \frac{\lambda_{2,n} - \lambda_{3,n}}{\lambda_n - \lambda_{2,n}}, \quad n^* = \frac{\lambda_n - \lambda_{2,n}}{p^*(1 - p^*)}, \quad s^* = \lambda_n - n^* p^*.$$

The parameters for the approximation are taken as

$$n'' = \lfloor n^* \rfloor, \quad s'' = \lfloor s^* \rfloor, \quad p'' = \frac{n^* p^* + \{s^*\}}{n''},$$

where $\{s^*\}$ is the decimal part of s^*. Thus, the approximation to $\mathcal{L}(S_n)$ is given by

$$s'' + B_N(n'', p''). \tag{4}$$

The next approximations are obtained by using Krawtchhouk expansion of S_n.

2.4. Approximation IV

Let $g(x, n, p)$ be the p.f. (probability function) of the binomial distribution $B_N(n, p)$. The Krawtchouk polynomial $L_j^n(x, n, p)$ of degree j is defined by

$$L_j^n(x, n, p) = \frac{d^j}{dp^j} g(x, n, p) \big/ g(x, n, p).$$

Let Δ be the difference operator such that $\Delta^j g(x, n, p) = \Delta^{j-1} g(x-1, n, p) - \Delta^{j-1} g(x, n, p)$ $(j = 1, 2, \cdots)$ and $\Delta^0 g(x, n, p) = g(x, n, p)$. Using the operator Δ, the Krawtchouk polynomial $L_j^n(x, n, p)$ of degree j is written as

$$L_j^n(x, n, p) = n^{(j)} \Delta^j g(x, n-j, p) \big/ g(x, n, p).$$

The factorial moment generating function of S_n is $F(t) = E\big[(1+t)^{S_n}\big] = \prod_{j=1}^n (1 + p_j t)$ and its coefficient of t^m is equal to the factorial moment $E[S_n^{(m)}]$ of S_n, where $x^{(m)} = x(x-1)\cdots(x-m+1)$. Thus, we have

$$\mu_{(m)} = E[S_n^{(m)}] = \sum_{1 \le j_1 \ne \cdots \ne j_m \le n} p_{j_1} \cdots p_{j_m} \quad (m = 1, 2, \ldots, n),$$

where the summation of the right-hand side is taken over all integers $j_1, \ldots, j_m$ satisfying $1 \le j_1 \ne \cdots \ne j_m \le n$. Let $p_n(k) = \mu_{(k)}/n^{(k)}$, $p = p_n(1) = \bar{p}_n$ and

$$q_n(0) = 1, \quad q_n(j) = \sum_{i=0}^{j} (-1)^i \binom{j}{i} p^i p_n(j-i) \ \ (j = 1, 2, \ldots). \tag{5}$$

For example, $q_n(1) = 0$, $q_n(2) = p_n(2) - p^2$, $q_n(3) = p_n(3) - 3pp_n(2) + 2p^3$. Then, It holds that

$$P(S_n = x) = g(x, n, p) \left\{ 1 + \sum_{j=2}^{n} \frac{q_n(j)}{j!} L_j^n(x, n, p) \right\}.$$

These results are derived without the assumption of independence of $X_1, X_2, \ldots, X_n$ (Takeuchi and Takemura (1987)). Using the difference operator, $P(S_n = x)$ is also expressed as

$$P(S_n = x) = g(x, n, p) + \sum_{j=2}^{n} \binom{n}{j} q_n(j) \Delta^j g(x, n-j, p). \tag{6}$$

Under the independence of $X_1, X_2, \ldots, X_n$ assumed in this paper, Roos (2006) gives the following expression, which is called the Krawtchouk expression by him:

$$P(S_n = x) = \sum_{j=0}^{n} a_j(p) \Delta^j g(x, n-j, p), \tag{7}$$

where $a_0(p) = 1$ and

$$a_j(p) = \sum_{1 \le k_1 < \cdots < k_j \le n} \prod_{r=1}^{j} (p_{k_r} - p) \quad (j = 1, \ldots, n).$$

Since we put $p = p_n(1) = \bar{p}_n$, we have $a_1(p) = 0$. From (5), we have

$$\binom{n}{j} q_n(j) = \sum_{i=0}^{j} \binom{n-i}{j-i} \frac{1}{i!} (-p)^{j-1} \mu_{(i)},$$

which is equal to $a_j(p)$, by Roos (2000; p.4,(11)). Therefore, the expression (6) is identical to (7). We note that Krawtchouk polynomial $K_j^n(x, n, p)$ defined by Roos (2000; p.3, (8)) is equal to $p^j(1-p)^j L_j^n(x, n, p)$.

For $s = 0, 1, \ldots, n$, let $\mathcal{B}_s(x, n, p)$ be the finite signed measure such that

$$\mathcal{B}_s(x, n, p) = \sum_{j=0}^{s} a_j(p) \Delta^j g(x, n-j, p) \quad (x = 0, 1, \ldots, n). \tag{8}$$

Then for $s = 0, 1, \ldots, n$ and $j = 0, 1, \ldots, s$,

$$\sum_{x=j}^{n} x^{(j)} \mathcal{B}_s(x, n, p) = \mu_{(j)},$$

which means that the first s moments of $\mathcal{B}_s$ are equal to that of S_n (Roos(2000)). This relation is also proved by the methods similar to the fourth and third lines from the bottom of Takeuchi and Takemura (1987, p.90).

Based on (8), we consider the three approximations to $\mathcal{L}(S_n)$. The first one is $\mathcal{B}_0(x, n, \bar{p}_n) = \mathcal{B}_1(x, n, \bar{p}_n) = g(x, n, \bar{p}_n)$ which is equal to the approximation I. The second is $\mathcal{B}_2(x, n, \bar{p}_n)$ which is given by Takeuchi (1975; p.84). The third is $\mathcal{B}_3(x, n, \bar{p}_n)$. Let $\gamma_k(p) = \sum_{j=1}^{n} (p_j - p)^k$ $(k = 1, 2, 3)$. Since $p = \bar{p}_n$, we have $a_1(p) = \gamma_1(p) = 0$, $a_2(\bar{p}_n) = -\gamma_2(\bar{p}_n)/2$ and $a_3(\bar{p}_n) = \gamma_3(\bar{p}_n)/3$. Then, the second and third approximations are written as follows:

$$\mathcal{B}_2(x, n, \bar{p}_n) = g(x, n, \bar{p}_n) - \frac{\gamma_2(\bar{p}_n)}{2} \Delta^2 g(x, n-2, \bar{p}_n), \tag{9}$$

and

$$\mathcal{B}_3(x, n, \bar{p}_n) = g(x, n, \bar{p}_n) - \frac{\gamma_2(\bar{p}_n)}{2} \Delta^2 g(x, n-2, \bar{p}_n) + \frac{\gamma_3(\bar{p}_n)}{3} \Delta^3 g(x, n-3, \bar{p}_n), \tag{10}$$

where

$$\Delta^2 g(x, n-2, p) = \frac{g(x, n, p)}{n(n-1)p^2(1-p)^2} \Big\{ x^2 - \big[1 + 2(n-1)p\big] x + n(n-1)p^2 \Big\},$$

$$\begin{aligned} \Delta^3 g(x, n-3, p) = \frac{g(x, n, p)}{n^{(3)} p^3 (1-p)^3} \Big\{ & x^3 - 3\big[(n-2)p + 1\big] x^2 \\ & + \big[3(n-1)(n-2)p^2 + 3(n-2)p + 2\big] x - n(n-1)(n-2)p^3 \Big\}. \end{aligned}$$

$\mathcal{B}_2(x, n, \bar{p}_n)$ has the first two moments as same as S_n and $\mathcal{B}_3(x, n, \bar{p}_n)$ has the first three moments as same as S_n.

3. Ewens sampling formula

In this section, we consider the approximations to the distribution $\mathcal{L}(K_n)$ of the number K_n of distict components, for the Ewens sampling formula. Let the random variables $\xi_1, \xi_2, \cdots$ be independent and $P(\xi_j = 1) = p_j$, $P(\xi_j = 0) = 1 - p_j$ $(j = 1, 2, \ldots)$, where

$$p_j = \frac{\theta}{\theta + j - 1}, \quad (j = 1, 2, \ldots;\ \theta > 0).$$

Then the number K_n can be expressed as $K_n = \xi_1 + \xi_2 + \cdots + \xi_n$ $(n = 1, 2, \ldots)$. Since $\xi_1 = 1$ a.s. (almost surely), K_n can be expressed as

$$K_n = 1 + L_n \ \text{ a.s.}, \tag{11}$$

where $L_n = \xi_2 + \cdots + \xi_n$ $(n = 2, 3, \ldots)$. We derive the binomial approximation to $\mathcal{L}(L_n)$ using the results of Section 2 with $n - 1$. Since $p_j = \theta/(\theta + j - 1)$, we have

$$\lambda_{n-1} = \sum_{i=2}^{n} \frac{\theta}{\theta + i - 1} = \theta[\psi(\theta + n) - \psi(\theta + 1)],$$

$$\lambda_{2,n-1} = \sum_{i=2}^{n} \Big(\frac{\theta}{\theta + i - 1}\Big)^2 = \theta^2[\psi'(\theta + 1) - \psi'(\theta + n)],$$

and

$$\lambda_{3,n-1} = \sum_{i=2}^{n} \Big(\frac{\theta}{\theta + i - 1}\Big)^3 = \theta^3[\psi''(\theta + n) - \psi''(\theta + 1)],$$

where ψ, ψ' and ψ'' are the digamma, trigamma and tetragamma functions, respectively. From λ_{n-1}, $\lambda_{2,n-1}$ and $\lambda_{3,n-1}$, we can calculate

$$\bar{p}_{n-1} = \frac{\lambda_{n-1}}{n-1}, \quad \gamma_2(\bar{p}_{n-1}) = \lambda_{2,n-1} - (n-1)\bar{p}_{n-1}^2,$$

and

$$\gamma_3(\bar{p}_{n-1}) = \lambda_{3,n-1} - 3\lambda_{2,n-1}\bar{p}_{n-1} + 2(n-1)\bar{p}_{n-1}^3.$$

Using these values, we can get Approximations I, II, III and IV to $\mathcal{L}(L_n)$. Shifting these approximations to the right by 1, we can obtain the approximations to $\mathcal{L}(K_n)$. For Approximations I, II, III and IV of Section 2, we replace n and λ_n, $\lambda_{2,n}$, $\lambda_{3,n}$ by $n - 1$ and the above λ_{n-1}, $\lambda_{2,n-1}$, $\lambda_{3,n-1}$, respectively. Then, we can write the approximations to $\mathcal{L}(K_n)$ as follows:

$$\text{Approx. I} \ : \ 1 + B_N(n-1, \bar{p}_{n-1}), \qquad \text{Approx. II} \ : \ 1 + B_N((n-1)', p'),$$

$$\text{Approx. III} \ : \ 1 + s'' + B_N((n-1)'', p''),$$

and

$$\text{Approx. IV.1} \ : \ 1 + \mathcal{B}_2(x, n-1, \bar{p}_{n-1}), \ \ \text{IV.2} \ : \ 1 + \mathcal{B}_3(x, n-1, \bar{p}_{n-1}).$$

For $n = 25, 50$ and $\theta = 0.5, 10$, we illustrate Approximations I, II, III, IV.1 and IV.2 to $\mathcal{L}(K_n)$. The p.f. of K_n are drawn by the simulation using R, as bar graph.

Figures 1 and 3 are for $n = 25$ and $\theta = 0.5, 10$. Approximations I, II and III are drawn by the dashed, solid and dotted lines, respectively. Figures 2 and 4 are for $n = 25$

and $\theta = 0.5, 10$. Approximations I, IV.1 and IV.2 are drawn by the dashed, solid and dotted lines, respectively.

Figures 5 and 7 are for $n = 50$ and $\theta = 0.5, 10$. Approximations I, II and III are drawn by the dashed, solid and dotted lines, respectively. Figures 6 and 8 are for $n = 50$ and $\theta = 0.5, 10$. Approximations I, IV.1 and IV.2 are drawn by the dashed, solid and dotted lines, respectively.

For the small θ, for example $\theta = 0.5$, Figures 1, 2, 5 and 6 show that there are little difference among Approximations I, II, III, IV.1 and IV.2. Figures 1, 3, 5 and 7 show that there are little difference between Approximations II and III. From Figures 2, 4, 6 and 8, it is seen that there are little difference between Approximations IV.1 and IV.2.

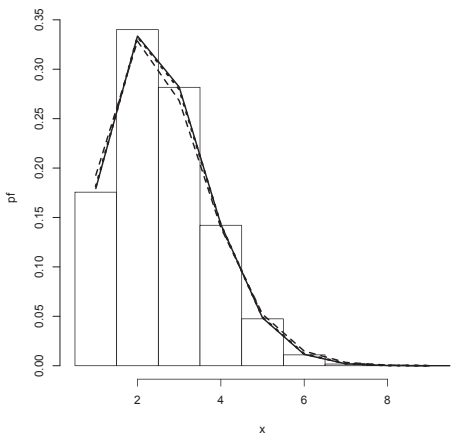

Fig. 1: I (dash), II (solid), III (dot) ($n = 25$, $\theta = 0.5$)

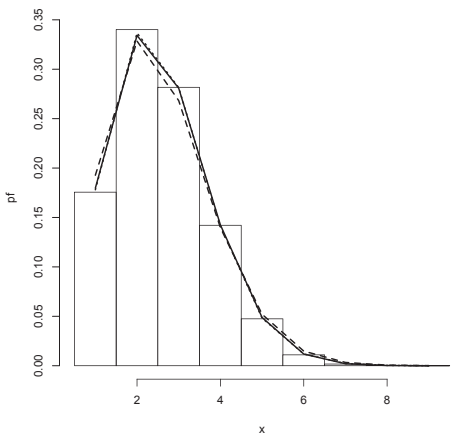

Fig. 2: I (dash), IV.1 (sol), IV.2 (dot) ($n = 25$, $\theta = 0.5$)

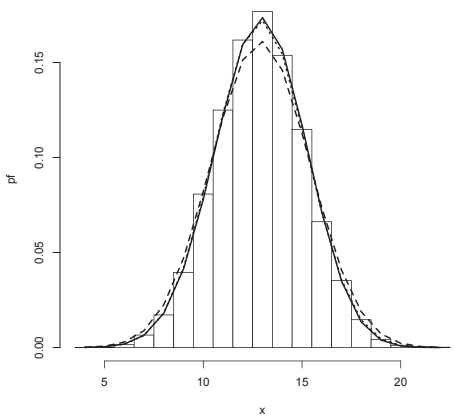

Fig. 3: I (dash), II (solid), III (dot) ($n = 25$, $\theta = 10$)

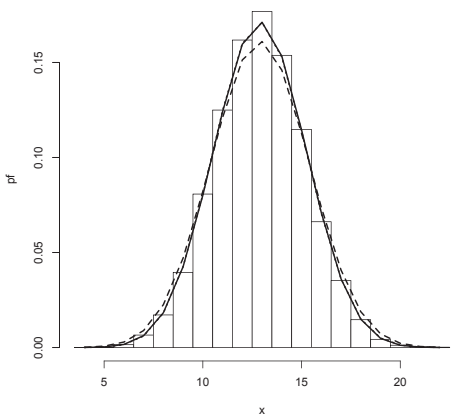

Fig. 4: I (dash), IV.1 (sol), IV.2 (dot) ($n = 25$, $\theta = 10$)

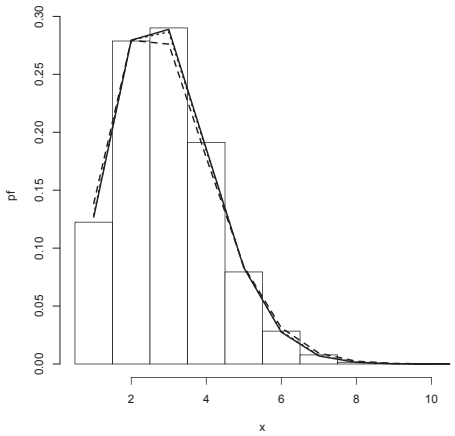

Fig. 5: I (dash), II (solid), III (dot) ($n = 50$, $\theta = 0.5$)

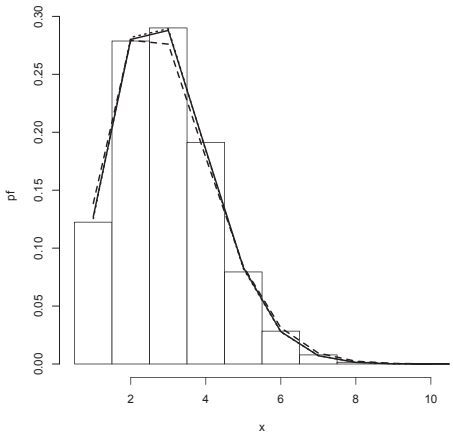

Fig. 6: I (dash), IV.1 (sol), IV.2 (dot) ($n = 50$, $\theta = 0.5$)

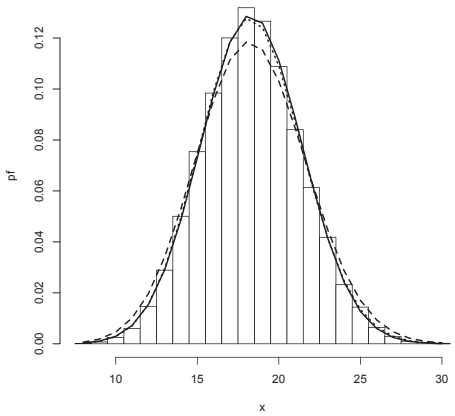

Fig. 7: I (dash), II (solid), III (dot) ($n = 50$, $\theta = 10$)

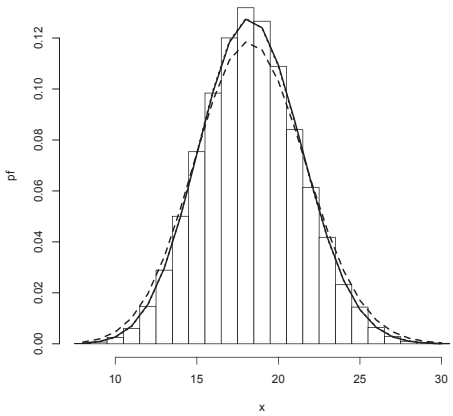

Fig. 8: I (dash), IV.1 (sol), IV.2 (dot) ($n = 50$, $\theta = 10$)

From all figures, it is seen that there are little difference between Approximations II and IV.1. While these two have the (almost same) first two moments, they are obtained by the distinct methods. We consider, as the approximation to the distribution of K_n, Approximations II and IV.1 are preferable. From the attached figures, we can not see the detailed differences among these approximations. If the reader wants the original pdf figures, please contact the author (yamato_march@hiz.bbiq.jp).

4. Discussion

Based on the approximate distribution to the sum of independent Bernoulli random variables, we give the shifted binomial approximations to $\mathcal{L}(K_n)$ for the Ewens sampling formula and illustrate them. It is a future problem to investigate the properties of these shifted binomial approximations, especially Approximations II and IV.1.

Acknowledgement

This work was supported by Grant-in-Aid for Scientific Research (B) (No. 16H02791), Japan Society for the Promotion of Science.

References

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics*, **2**, 1152–1174.

Arratia, R., Barbour, A.D. and Tavaré, S. (2000). The number of components in logarithmic combinatorial structure. *Annals of Applied Probability*, **10**, 331–361.

Arratia, R., Barbour, A.D. and Tavaré, S. (2003). Logarithmic combinatorial structures: a probabilistic approach. EMS Monographs in Mathematics, EMS Publishing House, Zürich.

Barbour, A.D., Holst, L. and Janson, S. (1992). Poisson approximation. Clarendon Press, Oxford.

Čekanavičius, Peköz, E., Röllin, A. and Shwartz, M. (2009). A three-parameter binomial approximation. *Journal of Applied Probobability*, **46**, 1073–1085

Ehm, W (1991). Binomial approximation to the Poisson binomial distribution. *Statistics & Probability Letters*, **11**, 7–16

Ewens, W.J. (1972). The sampling theory of selectively neutral alleles. *Theoretical population biology*, **3**, 87–112.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, **1**, 209–230.

Johnson, N.L., Kotz, S. and Balakrishnan, N. (1997). Discrete multivariate distributions. John Wiley & Sons, New York.

Roos, B. (2000). Binomial approximation to the Poisson binomial distribution: The Krawtchouk expansion. *Theory of probability and its applications*, **45**, 258–272.

Soon, S.Y.T. (1996). Binomial approximation for dependent indicators. *Statistica Sinica*, **6**, 703–714.

Takeuchi, K. (1975). *Approximation of probability distributions* (in Japanese). Kyouiku Shuppan, Tokyo.

Takeuchi, K. and Takemura, A. (1987). On sum of 0-1 random variables I. univariate case. *Ann. Inst. Statist. Math.*, **39**, 85–102.

Yamato, H. (2017). Poisson approximations for sum of bernoulli random variables and its application to Ewens sampling formula. *J. Japan Statist. Soc.*, to appear.

Yamato, H., Nomachi, T. and Toda, K. (2015). Approximate distributions for the number of distinct components of the Ewens sampling formula and its applications. *Bulletin of Informatics and Cybernetics*, **47**, 69–81.

Received March 21, 2017
Revised October 23, 2017