APPROXIMATE DISTRIBUTIONS FOR THE NUMBER OF DISTINCT COMPONENTS OF THE EWENS SAMPLING FORMULA AND ITS APPLICATIONS

Yamato, Hajime
Kagoshima University : Emeritus

Nomachi, Toshifumi
Miyakonojo College of Technology

Toda, Koichiro
Kawai-juku prep school

https://doi.org/10.5109/1909525

出版情報：Bulletin of informatics and cybernetics. 47, pp.69-81, 2015-12. Research Association of Statistical Sciences
バージョン：
権利関係：

Bulletin of Informatics and Cybernetics, Vol. 47, 2015

APPROXIMATE DISTRIBUTIONS FOR THE NUMBER OF DISTINCT COMPONENTS OF THE EWENS SAMPLING FORMULA AND ITS APPLICATIONS

By

Hajime Yamato[*], **Toshifumi Nomachi**[†] and **Koichiro Toda**[‡]

Abstract

The Ewens sampling formula is well-known as a distribution of a random partition of a positive integer n or a set of integers $\{1, 2, \cdots, n\}$. The number K_n of distinct components of the formula has the asymptotic normality. For its well-known form and the related form, Yamato(2013) gives their Edgeworth expansions. But, their appropriateness depend on the parameter. Using the functions of R, we consider its normal approximation suitable for any value of the parameter. As the application, we show the method to search the the maximum likelihood estimator of the parameter graphically, and gives its approximate distribution. We also consider the approximate distribution of K_n in case where the parameter of the formula is the random variable. These results are shown with the graphs.

Key Words and Phrases: Ewens sampling formula, maximum likelihood estimator, mixing distribution, normal approximation, random partition.

1. Introduction

Ewens (1972) discovered a distribution of a random partition of a positive integer n or a set of integers $\{1, 2, \cdots, n\}$, 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 context. For example, Bayesian statistics, pattern of communication and genetics. There are many works on the Ewens sampling formula and the related formula, which includes Pitman's sampling formula (Pitman (1992)). See, for example, Johnson et al. (1997), Arratia et al. (2003) and Pitman (2006).

On the other hand, in case where the parameter of the Ewens sampling formula is the random variable having the distribution G, the distribution of a random partition of the positive integer is given by mixing the Ewens sampling formula with G (see, Gnedin and Pitman (2006)). It was also derived by the mixture of Dirichlet process (see, for the mixture of Dirichlet process, Antoniak (1974)).

In this paper, at first, we consider the Ewens sampling formula and let K_n be the number of distinct components of the sampling formula with the parameter $\theta > 0$.

[*] Emeritus of Kagoshima University, Take 3-32-1-708, Kagoshima 890–0045, Japan
[†] Miyakonojo College of Technology, Miyakonojo, Miyazaki 885–8567, Japan
[‡] Kawai-juku prep school, Fukuoka 810–8619, Japan

Let $\{\xi_j, j = 1, 2, \cdots\}$ be a sequence of independent random variables whose probability functions are given by $P(\xi_j = 0) = (j-1)/(\theta + j - 1)$ and $P(\xi_j = 1) = \theta/(\theta + j - 1)$ $(j = 1, 2, \cdots)$. Then K_n can be written as

$$K_n = \xi_1 + \xi_2 + \cdots + \xi_n \quad (n = 1, 2, \cdots). \tag{1}$$

The distribution of K_n is given by

$$P(K_n = k) = \left[\begin{array}{c} n \\ k \end{array} \right] \frac{\theta^k}{\theta^{[n]}} \quad (k = 1, 2, \ldots, n), \tag{2}$$

where $\theta^{[n]} = \theta(\theta+1)\cdots(\theta+n-1)$ and $\left[\begin{array}{c} n \\ k \end{array} \right]$ is a signless Stirling number of the first kind or a Stirling number of the third kind. The mean and variance of K_n are

$$E(K_n) = \theta \sum_{i=1}^{n} \frac{1}{\theta + i - 1}, \quad V(K_n) = \theta \sum_{j=1}^{n} \frac{1}{\theta + j - 1} - \theta^2 \sum_{j=1}^{n} \frac{1}{(\theta + j - 1)^2}. \tag{3}$$

K_n has asymptotically the normal distribution $N(\theta \log n, \theta \log n)$. (See, for example, Johnson et al (1997; Chapter 41) and Arratia et al. (2003; Section 5.2)). Yamato (2013) gives the following Edgeworth expansions of K_n, which hold uniformly in $x \in \mathbb{R}$.

$$P\left(\frac{K_n - \theta \log n}{\sqrt{\theta \log n}} \le x \right) = \Phi(x) - \frac{1}{6\sqrt{\theta \log n}} \phi(x)\big(x^2 - 1 - 6\theta\psi(\theta)\big) + O\left(\frac{1}{\log n} \right), \tag{4}$$

and

$$P\left(\frac{K_n - \theta(\log n - \psi(\theta))}{\sqrt{\theta(\log n - \psi(\theta))}} \le x \right) = \Phi(x) - \frac{1}{6\sqrt{\theta \log n}} \phi(x)(x^2 - 1) + O\left(\frac{1}{\log n} \right), \tag{5}$$

where ψ is the digamma function defined by $\psi(\theta) = \Gamma'(\theta)/\Gamma(\theta)$, and Φ and ϕ are the distribution function and the density function of $N(0, 1)$, respectively.

On the other hand, in case where the parameter θ of the Ewens sampling formula is the random variable having the distribution G, Yamato (2012) shows that $K_n/\log n$ converges to G in distribution as $n \to \infty$. Yamato and Kondo (2014) gives its Edgeworth expansions. Under the suitable conditions for the density function g of G, for example, it is given by

$$P\left(\frac{K_n}{\log n} \le x \right) = G(x) + \frac{1}{2 \log n} \Big\{ \big(2x\psi(x+1) - 1\big) g(x) + x g'(x) \Big\} + O\left(\frac{1}{(\log n)^{2/5}} \right). \tag{6}$$

In the section 2.1, as the approximate distribution of K_n of the Ewens sampling formula, we illustrate the distributions based on (4), (5) and the related distribution, graphically. The pictures show that these distributions are not over all good approximation to the distribution of K_n. The local limit evaluation of K_n which is given by Arratia et al. (2003;p.101,(5.20)) is also not over all good approximation. Thus we quit searching the simple form of the approximate distribution of K_n. In the section 2.2, we use the normal approximation of K_n with the mean and variance given by (3) as they stand. Because the mean and variance given by (3) can be calculated by R, we can easily draw the exact normal approximation of K_n which is over all good approximation

to the distribution of K_n. Since the mean of K_n can be drawn by R, in the section 2.3, we show that we can easily obtain the maximum likelihood estimate (MLE) of θ by the graph. The approximate distribution of MLE is also given.

In the section 3.1, we consider the case where the parameter θ of the Ewens sampling formula is the random variable. As the approximate distribution of K_n, we illustrate the distributions based on (6), which is not over all good approximation to the distribution of K_n. We show that the mixing distribution is over all good approximation to the distribution of K_n. The parameter of the mixing distribution is estimated graphically, which is shown in the section 3.2.

2. The approximate distributions of K_n

2.1. On the approximate distributions based on the asymptotic properties

At first, we consider the approximate distributions of K_n based on the relation (4) and illustrate

$$P(K_n \le x) \doteqdot \Phi(t) - \frac{1}{6\sqrt{\theta \log n}}\phi(t)\big(t^2 - 1 - 6\theta\psi(\theta)\big), \quad t = \frac{x - \theta \log n}{\sqrt{\theta \log n}} \tag{7}$$

for $n = 50$ and $\theta = 1, 2, 5$. The distribution function of K_n is simulated with R using (1), and drawn by the step function. The normal distribution function of $N(\theta \log n, \theta \log n)$ and the Edgeworth expansion (7) are drawn by the dashed curve and the dotted one, respectively. The approximate distributions of K_n based on (4) are good for the small θ such as $\theta < 2$, as shown by Figures 1,2. As the Figure 3 shows, the approximate distributions of K_n based on (4) are not appropriate for the large such as $\theta \ge 5$.

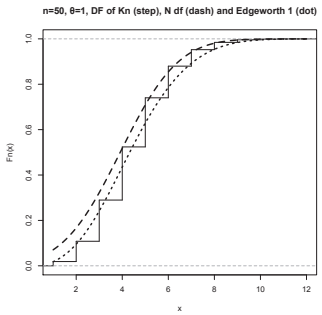

Fig. 1: $n = 50, \theta = 1$

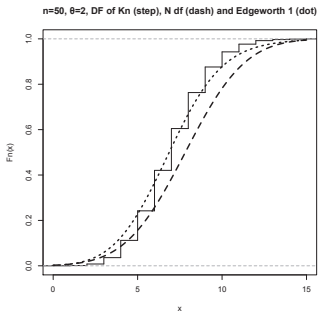

Fig. 2: $n = 50, \theta = 2$

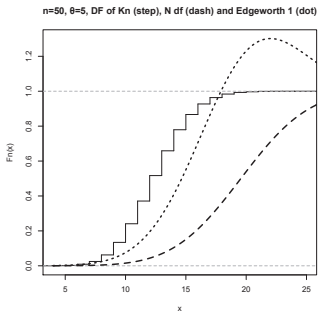

Fig. 3: $n = 50, \theta = 5$

Next, we consider the approximate distributions of K_n based on the relation (5) and illustrate

$$P(K_n \le x) \doteqdot \Phi(t) - \frac{1}{6\sqrt{\theta \log n}}\phi(t)(t^2 - 1), \quad t = \frac{x - \theta(\log n - \psi(\theta))}{\sqrt{\theta(\log n - \psi(\theta))}} \tag{8}$$

for $n = 50$ and $\theta = 2, 5, 15$. The distribution function of K_n is simulated with R using (1), and drawn by the step function. The normal distribution function of $N(\theta(\log n - \psi(\theta)), \theta(\log n - \psi(\theta)))$ and the Edgeworth expansion (8) are drawn by the dashed curve and the dotted one, respectively. The approximate distributions of K_n based on (5) are

good for the small θ such as $\theta < 5$, as shown by Figures 4,5. As the Figure 6 shows, the approximate distributions of K_n based on (5) are not appropriate for the large such as $\theta \geq 15$. In this case, Figures 4,5,6 shows that the effect of the Edgeworth expansion which is the right-hand side of (8) is very little.

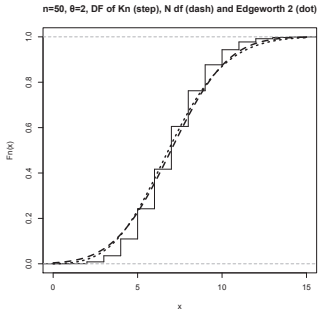

Fig. 4: $n = 50, \theta = 2$

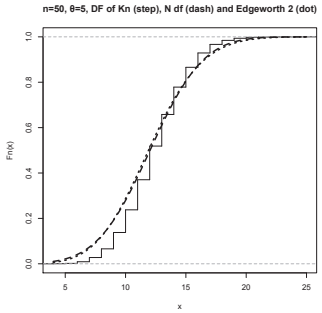

Fig. 5: $n = 50, \theta = 5$

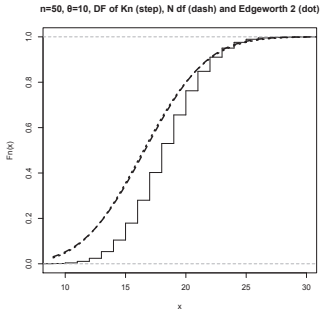

Fig. 6: $n = 50, \theta = 15$

These illustrations are based on the approximate distributions of K_n, given by $N(\theta \log n,\ \theta \log n)$ and $N(\theta(\log n - \psi(\theta)), \theta(\log n - \psi(\theta)))$. Alternatively, we consider their properties from the point of view of approximating the means $E(K_n)$ and the variances $V(K_n)$. For the mean, by the relation (2.5) of Yamato (2013) we have

$$\mid E(K_n) - \theta(\log n - \psi(\theta)) \mid = O\Big(\frac{1}{n}\Big). \tag{9}$$

For the variance, in addition, we use the the relation $\int_n^\infty (\theta + x)^{-2} dx = 1/(\theta + n) < \sum_{j=n}^\infty (\theta + j)^{-2} < \int_n^\infty (\theta + x - 1)^{-2} dx = 1/(\theta + n - 1)$ and $\sum_{j=1}^n (\theta + j - 1)^{-2} = \psi^{'}(\theta) - \sum_{j=n}^\infty (\theta + j)^{-2}$. Then, we have

$$\mid V(K_n) - \theta[\log n - \psi(\theta) - \theta\psi^{'}(\theta)] \mid = O\Big(\frac{1}{n}\Big) \tag{10}$$

where $\psi^{'}$ is the trigamma function.

For the mean associated with $n = 50, 100, 500$, the exact mean $E(K_n)$ is drawn by the solid curve, $\theta \log n$ by dashed curve, and $\theta(\log n - \psi(\theta))$ by dotted curve, as the function of θ. Figures 7,8,9 of the next page correspond to the cases of $n = 50, 100, 500$. $\theta \log n$ is good approximation of the mean $E(K_n)$ only for the θ close to zero. $\theta(\log n - \psi(\theta))$ is good approximation of the mean $E(K_n)$ even for the θ not close to zero.

For the variance associated with $n = 50, 100, 500$, the exact variance $V(K_n)$ is drawn by the solid curve, $\theta \log n$ by dashed curve, $\theta(\log n - \psi(\theta))$ by dotted curve, and $\theta[\log n - \psi(\theta) - \theta\psi^{'}(\theta)]$ by dot-dashed curve, as the function of θ. $\theta \log n$ is good approximation of the variance $V(K_n)$ only for the θ close to zero. $\theta(\log n - \psi(\theta))$ and $\theta[\log n - \psi(\theta) - \theta\psi^{'}(\theta)]$ are better approximations than $\theta \log n$, but the latter gives smaller value than the exact value. These are shown by Figures 10,11,12 of the next page, which correspond to the cases of $n = 50, 100, 500$, respectively.

Next, we consider the approximate distributions of K_n based on the approximate mean by (9) and the approximate variance by (10). That is,

$$K_n \dot{\sim} N\big(\theta(\log n - \psi(\theta)), \theta[\log n - \psi(\theta) - \theta\psi^{'}(\theta)]\big). \tag{11}$$

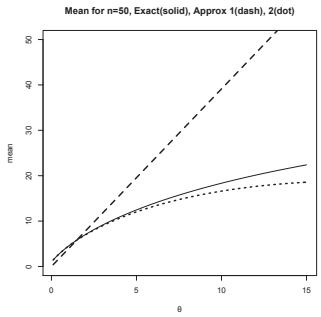

Fig. 7: for $n = 50$

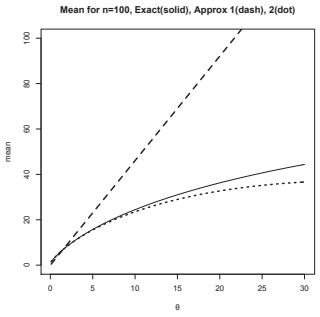

Fig. 8: Mean for $n = 100$

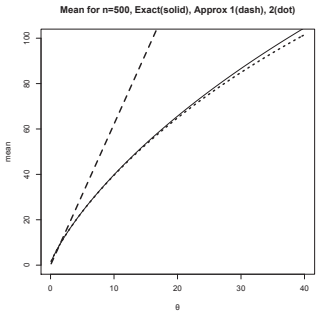

Fig. 9: Mean for $n = 500$

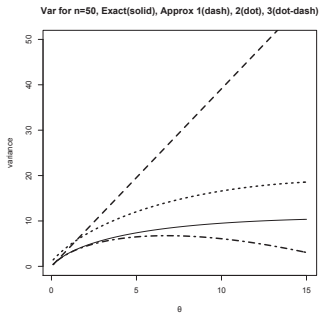

Fig. 10: Var for $n = 50$

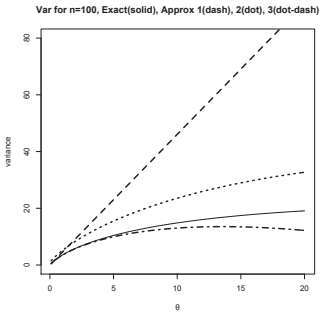

Fig. 11: Var for $n = 100$

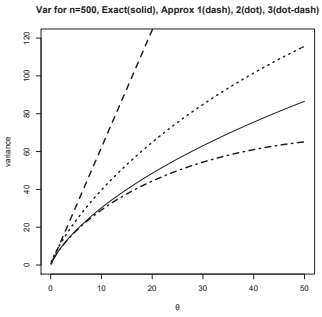

Fig. 12: Var for $n = 500$

For $n = 50$ and $\theta = 0.5, 2, 10$, the simulated distribution function of K_n is drawn by the step function. The normal distribution function (11) is drawn by the dotted curve. This approximate distribution is good for the small θ such as $\theta < 2$, as shown by Figures 13,14. As Figure 15 shows, the approximate distribution (11) are not appropriate for the large such as $\theta \geq 10$ and influenced by the under-estimate of the exact $V(K_n)$.

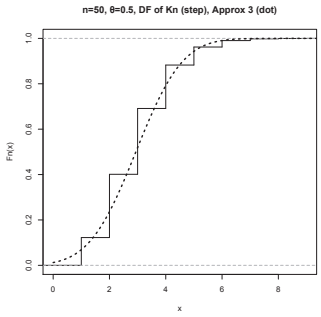

Fig. 13: $n = 50, \theta = 0.5$

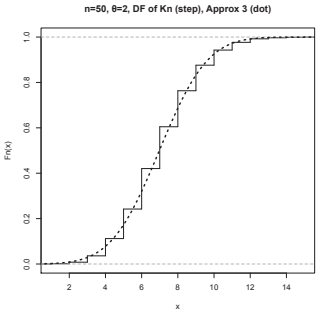

Fig. 14: $n = 50, \theta = 2$

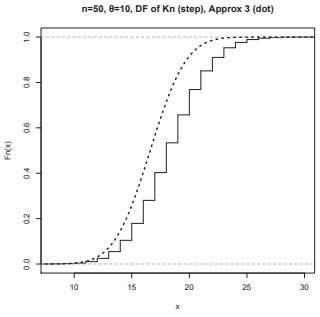

Fig. 15: $n = 50, \theta = 10$

At last, we consider the following local limit evaluation of K_n which is given by

Arratia et al. (2003;p.101,(5.20)). If $k/\log n \to \beta\theta$ and $\beta \in [0, \infty)$, it holds that

$$P(K_n = k) \sim \frac{(\theta \log n)^{k-1} e^{-\theta \log n}}{(k-1)!} \frac{\Gamma(\theta+1)}{\Gamma(\beta\theta+1)}.$$

In spite of the limiting operation such that $k/\log n \to \beta\theta$, we approximate the probability $P(K_n = k)$ by taking $k/\log n = \beta\theta$ and draw it with the dotted curve. The probability function of K_n is simulated with R, and drawn by the bar graph. The left-tail probability is the good approximation to the exact probability. But, except the left-tail, the probability is not good approximation because of the property of the limiting operation. For $n = 50$, these are shown by Figures 16, 17 ($\theta = 3$) and 18, 19 ($\theta = 4$).

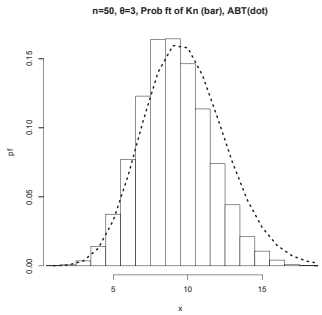

Fig. 16: $n = 50, \theta = 3$

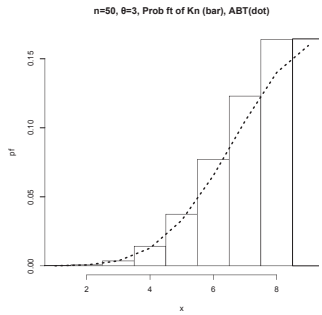

Fig. 17: $n = 50, \theta = 3$

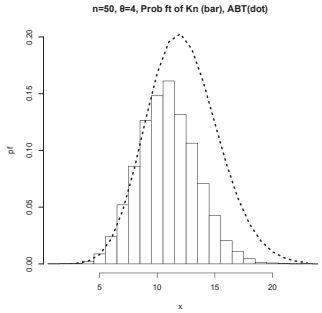

Fig. 18: $n = 50, \theta = 4$

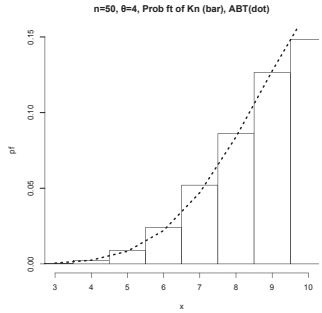

Fig. 19: $n = 50, \theta = 4$

In this section, we consider the approximate distributions of K_n by approximating the exact mean $\mu_n(\theta)$ and variance $\sigma_n^2(\theta)$, except the last local limit evaluation. All these approximate distributions do not give overall good approximation as shown. In the next section, we shall consider the approximate distribution of K_n which gives overall good approximation.

2.2. Approximate distribution of K_n using R

The digamma function ψ and trigamma function ψ' satisfy

$$\psi(x+n) - \psi(x) = \sum_{j=1}^{n} \frac{1}{x+j-1}, \quad \psi'(x+n) - \psi'(x) = -\sum_{j=1}^{n} \frac{1}{(x+j-1)^2},$$

respectively. Therefore, from (3) we have

$$\mu_n(\theta) = E(K_n) = \theta[\psi(\theta + n) - \psi(\theta)], \tag{12}$$

$$\sigma_n^2(\theta) = V(K_n) = \theta[\psi(\theta + n) - \psi(\theta)] + \theta^2[\psi^{'}(\theta + n) - \psi^{'}(\theta)]. \tag{13}$$

As stated in the last section, the approximate distributions of the section 2.1 do not give overall good approximation to the distribution of K_n. Since the digamma and trigamma functions are included in the free software R, as the approximate distribution of K_n we shall take $N(\mu_n(\theta), \sigma_n^2(\theta))$ with $\mu_n(\theta)$ and $\sigma_n^2(\theta)$ given by (12) and (13), respectively. This approximate distribution can be calculated and drawn by R.

For $n = 25$ and $\theta = 0.5, 1, 2, 5, 10, 15$, Figures 20, 21, 22, 23, 24, 25 give the distribution functions of K_n drawn by the step function, which are simulated with R. The normal distribution functions of $N(\mu_n(\theta), \sigma_n^2(\theta))$ are drawn by the dashed curves.

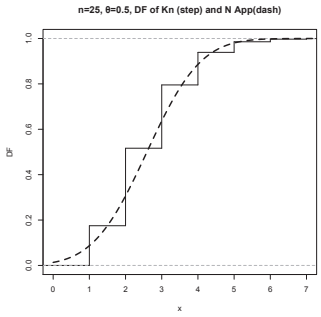

Fig. 20: $n = 25, \theta = 0.5$

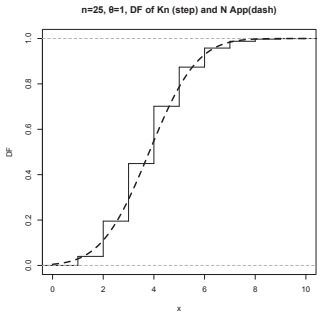

Fig. 21: $n = 25, \theta = 1$

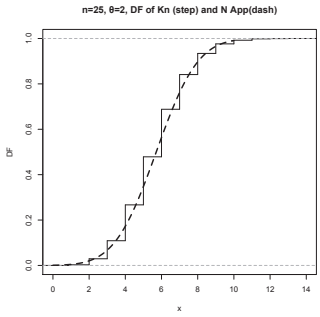

Fig. 22: $n = 25, \theta = 2$

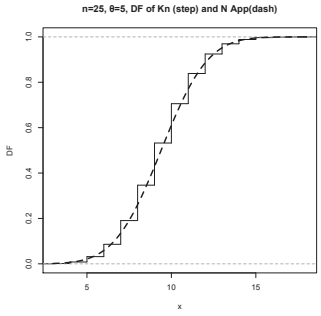

Fig. 23: $n = 25, \theta = 5$

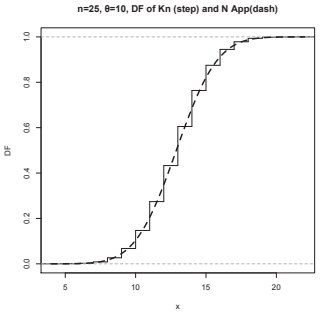

Fig. 24: $n = 25, \theta = 10$

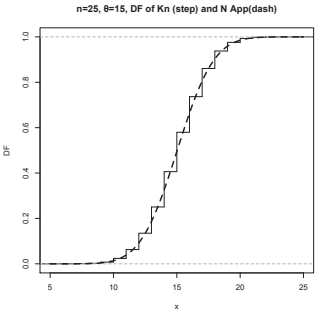

Fig. 25: $n = 25, \theta = 15$

For $n = 25$ and $\theta = 2, 5, 15$, Figures 26, 27, 28 of the next page give the simulated probability functions of K_n are drawn by the bar graph. The normal density functions of $N(\mu_n(\theta), \sigma_n^2(\theta))$ are drawn by the dashed curves.

These show that the normal distribution $N(\mu_n(\theta), \sigma_n^2(\theta))$ can be drawn by R and give overall good approximation to the distribution of K_n even for a small n such as 25.

2.3. Maximum likelihood estimator of θ

Differentiating the right-hand side of (2) by θ and putting it to zero, we obtain the likelihood equation $k = \sum_{j=1}^{n} \theta/(\theta + j - 1)$ which gives the maximum likelihood

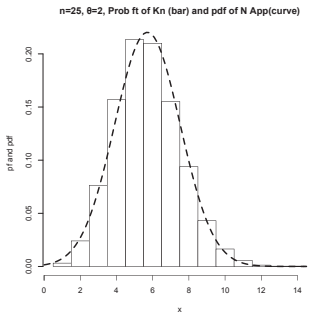

Fig. 26: $n = 25, \theta = 2$

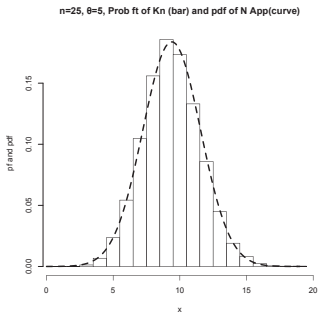

Fig. 27: $n = 25, \theta = 5$

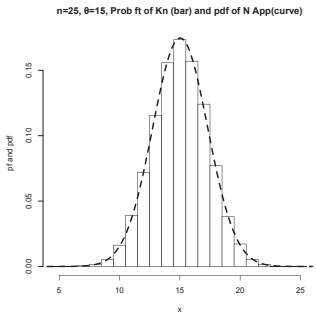

Fig. 28: $n = 25, \theta = 15$

estimator (MLE) of θ. The solutions of the equation are solved implicitly (see, for example, Ewens (1972; p.98) and Johnson et al. (1997; p.236)). We shall propose the method of obtaining the MLE of θ graphically.

Note that the likelihood equation is written as follows.

$$k = \mu_n(\theta), \quad \mu_n(\theta) = \theta[\psi(\theta + n) - \psi(\theta)]. \tag{14}$$

Since $d\mu_n(\theta)/d\theta = \sum_{j=1}^{n}(j-1)/(\theta + j - 1)^2 > 0$, the function $\mu_n(\theta)$ of θ is monotone increasing. Thus, from the graph of $\mu_n(\theta)$, we can read the MLE of θ for any given k.

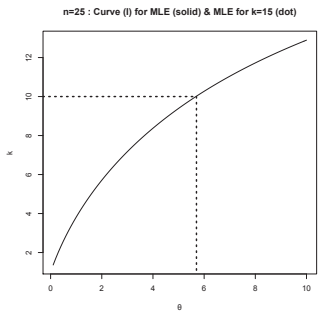

Fig. 29: $n = 25$

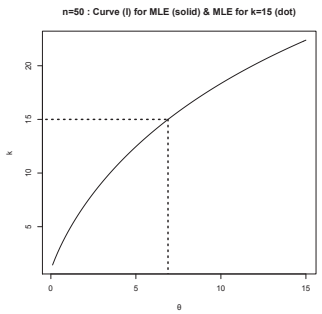

Fig. 30: $n = 50$

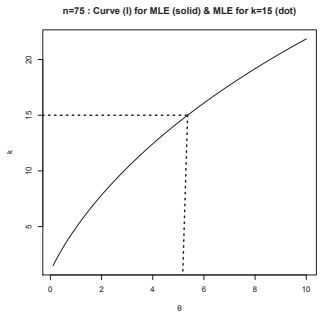

Fig. 31: $n = 75$

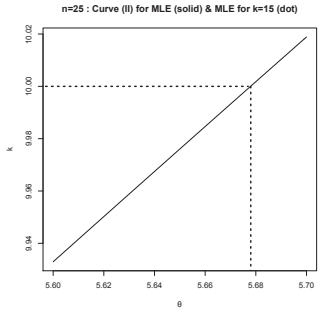

Fig. 32: $n = 25, k = 10$

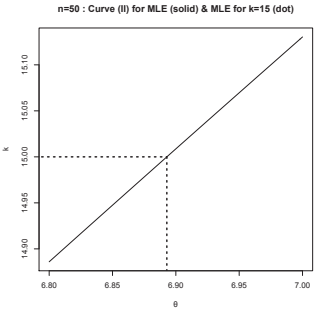

Fig. 33: $n = 50, k = 15$

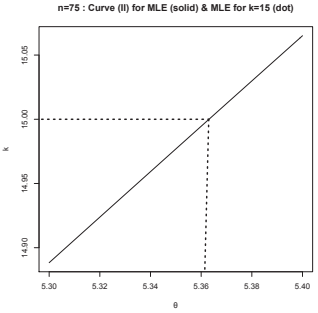

Fig. 34: $n = 75, k = 15$

Figures 29 and 32 are the curve of $\mu_n(\theta)$ in case of $n = 25$. For $k = 10$, we can read 5.7 as the MLE of θ. Figures 30 and 33 are the curve of $\mu_n(\theta)$ in case of $n = 50$. For $k = 15$, we can read 6.9 as the MLE of θ. Figures 31 and 34 are the curve of $\mu_n(\theta)$

in case of $n = 75$. For $k = 15$, we can read 5.4 as the MLE of θ.

Because of the monotone increasingness of $\mu_n(\theta)$, the value of K_n corresponds uniquely to the value of MLE $\hat{\theta}_n$ of θ. Therefore, for any value x of $\hat{\theta}_n$, the exact probability $P(\hat{\theta}_n = x)$ of $\hat{\theta}_n$ can be obtained from the probability function of K_n by the relation $P(\hat{\theta}_n = x) = P\big(\mu_n(\hat{\theta}_n) = \mu_n(x)\big) = P(K_n = \mu_n(x))$. On the other hand, from the approximate distribution $N(\mu_n(\theta), \sigma_n^2(\theta))$ of K_n, the approximate probability of $\hat{\theta}_n$ can be obtained. For any value x of $\hat{\theta}_n$, it is given by

$$P(\hat{\theta}_n = x) = P(K_n = \mu_n(x)) \doteqdot \frac{1}{\sigma_n(\theta)}\phi\left(\frac{\mu_n(x) - \mu_n(\theta)}{\sigma_n(\theta)}\right). \tag{15}$$

For $n = 50$ and $\theta = 1, 5, 10$, the simulated probability functions of MLE $\hat{\theta}_n$ are drawn by the bar in Figures 35,36,37 (compare with Tavaré (2004, Fig.3.2, p.42)). The dotted lines draw the approximate probabilities (15) of $\hat{\theta}_n$, which are good approximation.

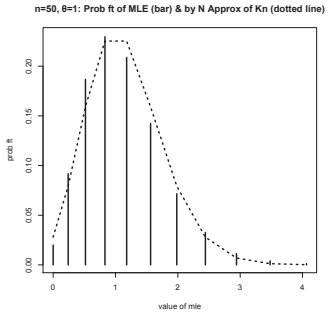

Fig. 35: $n = 50, \theta = 1$

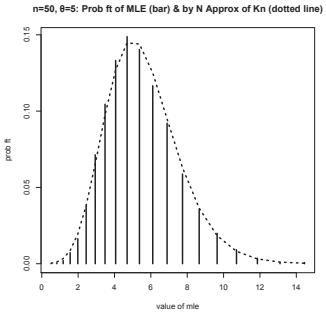

Fig. 36: $n = 50, \theta = 5$

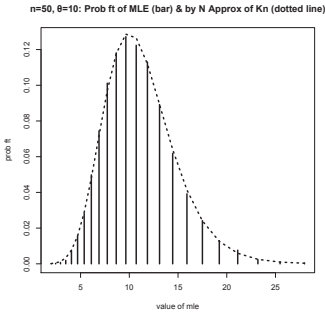

Fig. 37: $n = 50, \theta = 10$

The asymptotic distribution of MLE $\hat{\theta}_n$ is given by $N(\theta, \theta/\log n)$ (see, for example, Carlton (1999; p.80) and Tavaré (2004; p.41)). For $n = 50$ and $\theta = 5, 10$, Figures 38,39 show the density function of $N(\theta, \theta/\log n)$ drawn by the curves and the simulated probability functions of MLE $\hat{\theta}_n$ by the bar. These figures show that the asymptotic distribution $N(\theta, \theta/\log n)$ of MLE $\hat{\theta}_n$ concentrates around the true value of θ compared with the true distribution, for the small sample size. Thus, for the small sample size, $N(\theta, \theta/\log n)$ is not the good approximation to the distribution of MLE.

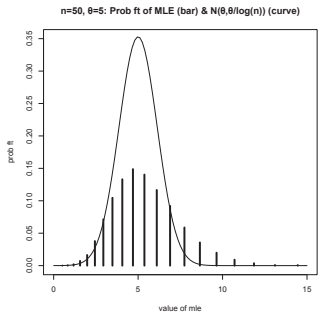

Fig. 38: $n = 50, \theta = 5$

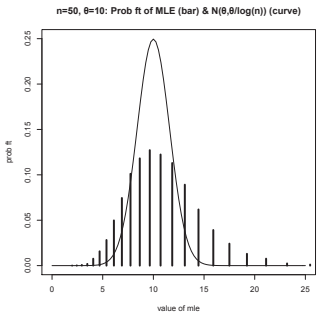

Fig. 39: $n = 50, \theta = 10$

3. The case where θ is a random variable

In this section, we suppose that $\theta(>0)$ is the random variable and has the Gamma distribution $Ga(c)$ whose density is given by $g_c(x) = x^{c-1}e^{-x}/\Gamma(c)$ $(c > 0)$.

3.1. Approximate distribution of K_n

We consider the case of $c > 1$ for the parameter of the Gamma distribution $Ga(c)$. Then from (6), we have the following, where G_c is the distribution function of $Ga(c)$.

$$P(K_n \leq x) \doteqdot G_c(t) + \frac{1}{2\log n}\Big\{\big(2t\psi(t+1)-1\big)g_c(t) + tg_c'(t)\Big\}, \quad t = \frac{x}{\log n}. \tag{16}$$

On the other hand, the normal distribution $N(\mu_n(\theta), \sigma_n^2(\theta))$ of the section 2.2 is the conditional distribution given θ under the situation of this section. Therefore, by taking the expectation of this conditional distribution $N(\mu_n(\theta), \sigma_n^2(\theta))$ with respect to the Gamma distribution $Ga(c)$, we have the mixture distribution as the approximate distribution of K_n. The followings $F_{K_n,c}(x)$ and $f_{K_n,c}(x)$ are the mixture distribution function and density function, respectively;

$$F_{K_n,c}(x) = \int_0^\infty \Phi\left(\frac{x-\mu_n(\theta)}{\sigma_n(\theta)}\right) dG_c(\theta) \tag{17}$$

and

$$f_{K_n,c}(x) = \int_0^\infty \frac{1}{\sigma_n(\theta)}\phi\left(\frac{x-\mu_n(\theta)}{\sigma_n(\theta)}\right) dG_c(\theta). \tag{18}$$

In the following, we consider the case of $n = 50$ and $c = 2, 4, 10$. The distributions of K_n are simulated by R using (1) and drawn with the step functions. The mixture distribution functions $F_{K_n,c}(x)$ are drawn by the dashed curves. The Edgeworth expansions given by the right-hand side of (16) and the Gamma distribution function G_c are drawn by the dotted curves and the dot-dashed ones, respectively (Figures 40, 41, 42). These pictures show that the mixture distribution functions $F_{K_n,c}(x)$ give overall good approximations to the distributions of K_n. Figures 43, 44, 45 of the next page show this fact by the probability functions of K_n drawn by the bar graph and the mixture density functions $f_{K_n,c}(x)$ by the dashed curves.

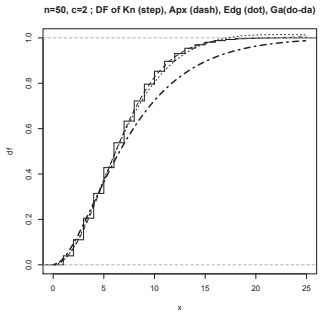

Fig. 40: $n = 50, c = 2$

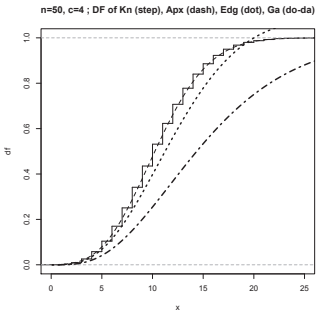

Fig. 41: $n = 50, c = 4$

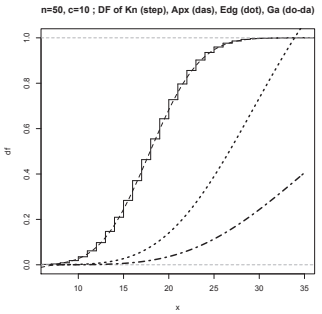

Fig. 42: $n = 50, c = 10$

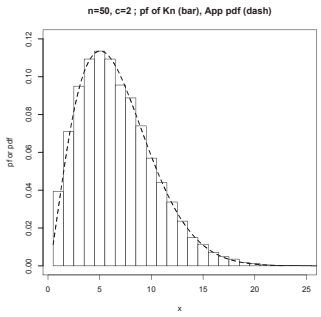

Fig. 43: $n = 50, c = 2$

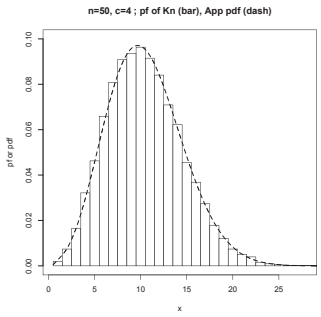

Fig. 44: $n = 50, c = 4$

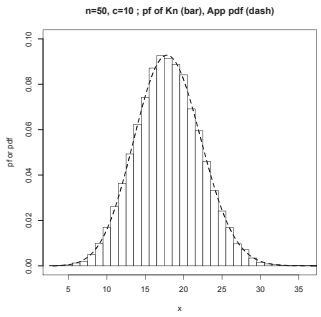

Fig. 45: $n = 50, c = 10$

3.2. Estimation of the parameter of the mixing distribution

We consider the estimation of the parameter c of the Gamma distribution $Ga(c)$. Since $E(K_n \mid \theta) = \mu_n(\theta)$ under the condition of this section because of (12), we have

$$E(K_n) = m_n(c), \quad m_n(c) = \int_0^\infty \mu_n(\theta) dG_c(\theta).$$

Therefore we can find the estimate of c by the method of moments, for given n and the observation k. It is enough to search c satisfying

$$k = m_n(c).$$

Since the gamma distribution G_c is stochastically increasing with respect to the parameter c and the function $\mu_n(\theta)$ $(\theta > 0)$ is positive and increasing, the right-hand side $m_n(c)$ is the increasing function of c. Thus, k of the left-hand side gives the unique solution c. By computing the integral $m_n(c)$ at the several points and drawing $m_n(c)$ by the smooth.spline of R, the unique solution c is easily read from the graph as Figures 46 and 47. From Figure 47, we get the estimate 4.95 of c for $n = 50$ and $k = 12$.

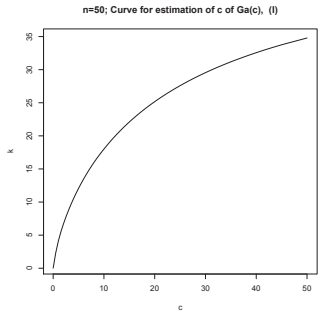

Fig. 46: $n = 50$

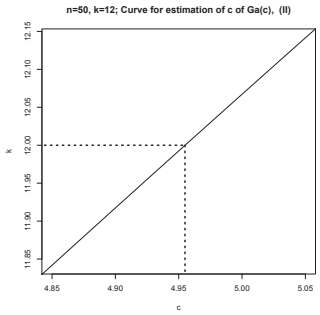

Fig. 47: $n = 50, k = 12$

4. Remark

Since the statistic K_n of the Ewens sampling formula can be written by the sum of the independent random variables as (1), it converges to the normal distribution as

$n \to \infty$. Because of the expressions (3) of the expectation $E(K_n)$ and the variance $V(K_n)$, the asymptotic normality of K_n is usually described by the normal distribution $N(\theta \log n, \theta \log n)$, as stated in the section 1. Yamato (2012) gives the another expressions of the asymptotic normality, which are $N(\theta(\log n - \psi(\theta)), \theta(\log n - \psi(\theta)))$. These asymptotic normal distributions and their Edgeworth expansions are not good approximations to the distribution of K_n for a large $\theta(> 0)$. Because, these asymptotic normal distribution are based on the approximations of $E(K_n)$ and $V(K_n)$, which are not good approximations for a large $\theta(> 0)$ as shown by Fig. 7,8,$\cdots$,12.

On the other hand, the expectation $E(K_n)$ and the variance $V(K_n)$ are written by the digamma and trigamma functions as (12) and (13). These functions are included in the free software R. Therefore, we can compute and draw the normal distribution $N(E(K_n), V(K_n))$, uing R. The reason why this normal distribution gives good approximation to the distribution of K_n for any $\theta(> 0)$ may be that the digamma and trigamma functions of R give good approximations to these functions. This result also gives the good approximation to the distribution of K_n, in case where the parameter θ of the Ewens sampling formula is the random variable.

Acknowledgement

The authors are grateful to the referee for his careful reading and useful comments.

References

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

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

Carlton, M. A. (1999). Applications of two-parameter Poisson-Dirichlet distribution. Ph.D. dissertation, University of California, L.A.

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.

Gnedin, A. and Pitman, J. (2006). Exchangeable Gibbs partitions and Stirling triangles. *J. of Mathematical Sciences*, **138**, Np.3, 5674–5685.

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

Pitman, J. (1992). The two-parameter generalization of Ewens random partition structure, Technical Report N. 345, Department of Statistics, University of California, Berkeley.

Pitman, J. (2006). *Combinatorial stochastic processes*, Springer, Berlin.

Tavaré, S. (2004). Ancestral inference in population genetics in *E*cole de'Eté Probabilités de Saint-Flour XXXI - 2001 of Lectures Notes in Mathematics, 3 – 188, Springer, Berlin.

Yamato, H. (2012). Asymptotic distribution of number of distinct observations among a sample from mixture of Dirichlet processes. *Bull. Inform. Cyber.*, **44**, 41–47.

Yamato, H. (2013). Edgeworth expansion for the number of distinct components associated with Ewens Sampling Formula. *J. Japan Statist. Soc*, **43**, 17–28.

Yamato, H. and Kondo, M. (2014). The Edgeworth expansion for the number of distinct observations with the mixture of Dirichlet processes. *Scientiae Mathematicae Japonicae*, **e-2014**, 297–304.

Received August 18, 2015
Revised November 27, 2015