SHIFTED BINOMIAL APPROXIMATION FOR THE EWENS SAMPLING FORMULA (II)

Yamato, Hajime
Kagoshima University : Professor emeritus

https://doi.org/10.5109/2233858

出版情報：Bulletin of informatics and cybernetics. 50, pp.43-50, 2018-12. Research Association of Statistical Sciences
バージョン：
権利関係：

Bulletin of Informatics and Cybernetics, Vol. 50, 2018

SHIFTED BINOMIAL APPROXIMATION FOR THE EWENS SAMPLING FORMULA (II)

By

Hajime Yamato*

Abstract

The Ewens sampling formula is well-known as the distribution of a random partition of the positive integer n into components. For the number K_n of distinct components of the formula, Yamato (2017a) gives the approximations to the distribution of K_n by using the shifted Binomial distributions and recommends the approximations II and IV.1 among them. We examine these two approximations furthermore, and compare them with the shifted Poisson approximation (Yamato (2017b)) and the Normal approximation (Yamato et al. (2015)). As applications of the approximation II, we give the two examples.

Key Words and Phrases: Approximate distribution, Binomial distribution, Ewens sampling formula, Normal distribution, Poisson distribution, Shifted distribution.

1. Introduction

Ewens (1972) discovered a distribution of a random partition of the positive integer n into components, 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 (1973)). The formula appears in many statistical contexts (see, for example, Johnson et al. (1997; Chap. 41) and Crane (2016)). For the Ewens sampling formula, the number K_n of distinct components has the distribution whose probability function is given by $P(K_n = k) = \mid s(n,k) \mid \theta^k/\theta^{[n]}$ $(k = 1, 2, \ldots, n)$, 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. 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 and trigamma functions and these functions are included in the programming language R, the Normal 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 distinct components is studied by Arratia et al. (2000) in detail with respect to the logarithmic combinatorial structure including the Ewens sampling formula. Differently from Arratia et al. (2000), Yamato (2017b) 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 Bernoulli random variables (for example, Barbour et al. (1992; p. 190) and Roos (2006)).

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

Using these results, Yamato (2017a) gives the approximations to $\mathcal{L}(K_n)$ by the shifted Binomial distributions. Among them, the author considered that the approximations II and IV.1 (following the notations of Yamato (2017a)) are preferable. Our purpose is to investigate these two approximations furthermore, and compare them with the Poisson and the Normal approximations. In Section 2, we quote the approximations II and IV.1 from Yamato (2017a) and examine them by the total variation distance. In Section 3, we compare the approximations II, IV.1, the shifted Poisson approximation, and the Normal approximation by illustration, using R. In Section 4, we give the two examples as applications of the approximations II.

2. Shifted Binomial approximations to the distribution of K_n

We consider the shifted Binomial approximations to the distribution $\mathcal{L}(K_n)$ of the number K_n of distinct components of the Ewens sampling formula. We use the same notations as Yamato (2017a). 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., K_n can be expressed as

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

where $L_n = \xi_2 + \cdots + \xi_n$ $(n = 2, 3, \ldots)$. We let

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

and

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

where ψ and ψ' are the digamma and trigamma functions, respectively.

We note that

$$E(L_n) = \lambda_{n-1}, \quad Var(L_n) = \lambda_{n-1}\Big(1 - \frac{\lambda_{2,n-1}}{\lambda_{n-1}}\Big). \tag{2}$$

If $\lambda_{2,n-1}/\lambda_{n-1}$ is small, then $Var(L_n)$ is close to $E(L_n)$ and therefore the Poisson distribution is appropriate for the approximation to $\mathcal{L}(L_n)$. In general, because of $E(L_n) > Var(L_n)$, the Binomial distribution may be appropriate for the approximation to $\mathcal{L}(L_n)$. By applying the Binomial approximations to $\mathcal{L}(L_n)$, we get the shifted Binomial approximations to $\mathcal{L}(K_n)$. We quote the two approximations II and V.1 from Yamato (2017a).

Approximation II:

The approximation II to $\mathcal{L}(K_n)$ is the shifted Binomial distribution given by

$$\text{II} : 1 + B_N((n-1)', p') \quad ((n-1)' = \lfloor \lambda_{n-1}^2/\lambda_{2,n-1} \rfloor \ \text{ and } \ p' = \lambda_{n-1}/(n-1)'), \tag{3}$$

where $\lfloor x \rfloor$ is an integer close to x. For $B_N((n-1)', p')$, see Barbour et al. (1992; p. 190).

Approximation IV.1:

We put $\bar{p}_{n-1} = \lambda_{n-1}/(n-1)$ and $\gamma_2(\bar{p}_{n-1}) = \lambda_{2,n-1} - (n-1)\bar{p}_{n-1}^2$.

Let $g_B(x; n, p)$ be the probability function of the Binomial distribution $B_N(n, p)$. Let Δ be the difference operator such that $\Delta^j g_B(x; n, p) = \Delta^{j-1} g_B(x-1; n, p) - \Delta^{j-1} g_B(x; n, p)$ $(j = 1, 2, \cdots)$ and $\Delta^0 g_B(x; n, p) = g_B(x; n, p)$. Let $\mathcal{B}_2$ be the finite signed measure such that

$$\mathcal{B}_2(n-1, \bar{p}_{n-1})(\{x\}) = g_B(x; n-1, \bar{p}_{n-1}) - \frac{\gamma_2(\bar{p}_{n-1})}{2} \Delta^2 g_B(x; n-3, \bar{p}_{n-1}),$$

where

$$\Delta^2 g_B(x; n-3, p) = \frac{g_B(x; n, p)}{(n-1)(n-2)p^2(1-p)^2} \Big\{ x^2 - \big[1 + 2(n-2)p\big] x + (n-1)(n-2)p^2 \Big\}.$$

The approximation IV.1 to $\mathcal{L}(K_n)$ is the shifted finite signed measure given by

$$\text{IV.1} \ : \ 1 + \mathcal{B}_2(n-1, \bar{p}_{n-1})(\{x\}). \tag{4}$$

For $\mathcal{B}_2(n-1, \bar{p}_{n-1})$, see Takeuchi (1975) and Roos (2006).

The mean of the approximation II is equal to $E(K_n)$ and its variance is approximately equal to $Var(K_n)$. The mean and variance of the Approximation IV.1 are equal to $E(K_n)$ and $Var(K_n)$, respectively. Since the Binomial distribution is determined by the mean and variance uniquely, it is inferred that the two approximations have the similar behavior. This fact is shown by the illustration (Yamato (2017a)). We show the difference between them by the total variation distance. We define the total variation distance d_{TV} between the signed measures Q_1 and Q_2 over $\{0, 1, 2, \cdots\}$ as follows;

$$d_{TV}(Q_1, Q_2) = \frac{1}{2} \sum_{j=0}^{\infty} \mid Q_1(j) - Q_2(j) \mid .$$

Here, we note that $\sum_{i=2}^{n} 1/(\theta + i - 1) \sim \log n$ and $\sum_{i=2}^{n} 1/(\theta + i - 1)^2 < \pi^2/6$. Then, by Barbour et al. (1992; p.190, (2.4)) we have

$$d_{TV}\big(\mathcal{L}(K_n), 1 + B_N((n-1)', p')\big) = O\big((\log n)^{-1}\big). \tag{5}$$

By Roos (2006; (20)), we have

$$d_{TV}\big(\mathcal{L}(K_n), 1 + \mathcal{B}_2(n-1, \bar{p}_{n-1})\big) = O\big((\log n)^{-3/2}\big). \tag{6}$$

By (5) and (6), as the approximation to $\mathcal{L}(L_n)$, the approximation IV.1 is better than the approximation II. But, the approximation IV.1 has the drawback such that it may have the negative tail, because it is obtained by using the first two terms of the expansion of $\mathcal{L}(K_n)$ based on the Krawtchouk polynomial. We show the negative tail of the approximation IV.1, by the examples. The figures 1,2,3,4 show the left tails for $\theta = 10$ and $\theta = 20$ and the right tails for $\theta = 0.5$ and $\theta = 1$, for $n = 50$. The approximation IV.1 is preferable to approximating the neighborhood of the center of $\mathcal{L}(L_n)$. As the approximation to $\mathcal{L}(L_n)$, we consider that the approximation II is better that the approximation IV.1.

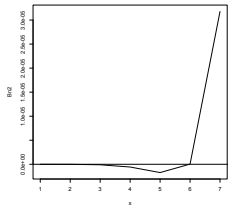

Fig. 1: $\theta = 10$, L

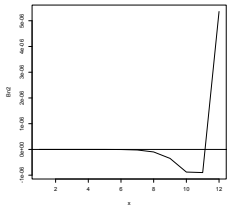

Fig. 2: $\theta = 20$, L

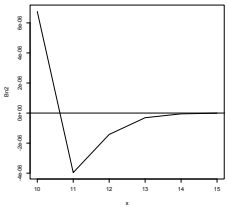

Fig. 3: $\theta = 0.5$, R

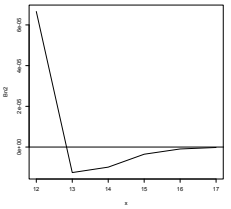

Fig. 4: $\theta = 1$, R

3. Shifted Binomial, Shifted Poisson, and Normal Approximations

Let $g_P(k;\mu)$ be the probability function of the Poisson distribution $Po(\mu)$. The shifted Poisson approximation to $\mathcal{L}(K_n)$ given by Yamato (2017b) is

$$\text{sPo} \ : \ 1 + g_P(k;\lambda_{n-1})\Big(1 - \frac{\lambda_{2,n-1}}{2} C_2(k,\lambda_{n-1})\Big), \tag{7}$$

where

$$C_2(x,\lambda) = \frac{x^2 - (2\lambda+1)x + \lambda^2}{\lambda^2}.$$

As the Normal approximation to $\mathcal{L}(K_n)$, we consider

$$\text{N} \ : \ N\,(\,E(K_n),\, Var(K_n)\,)) \tag{8}$$

(Yamato et al. (2015)), where $E(K_n) = \theta[\psi(\theta+n) - \psi(\theta)]$ and $Var(K_n) = \theta[\psi(\theta+n) - \psi(\theta)] + \theta^2[\psi'(\theta+n) - \psi'(\theta)]$.

By using R, we illustrate the comparison of the four approximations to $\mathcal{L}(K_n)$, which are the approximations II and IV.1 of the section 2, the shifted Poisson approximation (7), sPo and the Normal approximation (8), N. The probability function of K_n is simulated with R and drawn by the bar graph. In the following figures 5, 6, 7, 11, 12, 13, 17, 18, and 19, the Normal approximations are plotted by dashed lines and the approximation IV.1's are by dotted lines. In the figures 8, 9, 10, 14, 15, 16 , 20, 21, and 22, the shifted Poisson approximations are plotted by dashed lines and the approximation II's are by dotted lines. These figures show that the approximations II and IV.1 are good as the approximation to $\mathcal{L}(K_n)$.

Here, we note the Poisson distribution and the Normal distribution as the approximations to $\mathcal{L}(L_n)$ and $\mathcal{L}(K_n)$, respectively. Since

$$\lambda_{n-1} = \sum_{i=2}^{n} \frac{\theta}{\theta+i-1} > \theta[\log(\theta+n) - \log(\theta+1)],\ \lambda_{2,n-1} = \sum_{i=2}^{n} \Big(\frac{\theta}{\theta+i-1}\Big)^2 < \theta^2 \cdot \frac{\pi^2}{6},$$

we have

$$0 < \frac{\lambda_{2,n-1}}{\lambda_{n-1}} < \frac{\theta\pi^2}{6[\log(\theta+n) - \log(\theta+1)]}.$$

Therefore, if $\lambda_{2,n-1}/\lambda_{n-1}$ is small with a small θ or a large n ($\log n$), then the Poisson distribution is appropriate for the approximation to $\mathcal{L}(L_n)$ by (2). This is shown by the figures 8, 9 and 10 for $\theta = 0.125, 0.25$ and 0.5 and $n = 25$. For $\theta = 2$, the figures 14, 15 and 16 show that the shifted Poisson approximation gets better as n increases. But,

the figures 20, 21 and 22 show that a large n is necessary in order to obtain the good shifted Poisson approximation for a large θ.

On the other hand, we have

$$P(K_n = 1) = \frac{(n-1)!}{(\theta+1)^{[n-1]}} \uparrow 1 \text{ as } \theta \downarrow 0.$$

Thus, if $\theta(>0)$ is close to zero, then $P(K_n = 1)$ is close to 1 and the Normal distribution is not appropriate as the approximation to $\mathcal{L}(K_n)$. These facts are shown by the figures 5, 6, 7. For θ such as $P(K_n = 1)$ close to zero, the Normal approximation gets better as θ increases, which is shown by the figures 11, 12 and 13. If θ and n are large, then the Normal approximation is good, which is shown by the figures 17, 18 and 19.

4. Concluding Remarks

In conclusion, we recommend the approximation II as the approximation to $\mathcal{L}(K_n)$ among the four approximations II, VI.1, sPo, N. As the applications of the approximation II, we give the two examples **(i)** and **(ii)** as follows.

(i) The approximation to the probability function of MLE $\hat{\theta}$ of θ:
Given the observation $K_n = k$, the MLE $\hat{\theta}$ of the parameter θ is the solution of the equation

$$k = \sum_{j=1}^{n} \frac{\theta}{\theta + j - 1} \tag{9}$$

(Ewens (1972)). Using the digamma function ψ, (9) is written as

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

Since $\mu_n(\theta)$ is the strictly increasing function of θ, for each $k = 1, 2, \ldots, n$, there exists an unique $\mu_n^{-1}(k)$. Thus, we have

$$P(\hat{\theta} = \mu_n^{-1}(k)) = P(K_n = k) \quad (k = 1, 2, \ldots, n)$$

or

$$P(\hat{\theta} = x) = P(K_n = \mu_n(x)) \ (x = \mu_n^{-1}(k),\ k = 1, 2, \ldots, n).$$

Using the approximation II, the approximation to the probability function of MLE $\hat{\theta}$ is given by

$$P(\hat{\theta} = x) \fallingdotseq g_B(k-1; (n-1)', p') \quad (\mu_n(x) = k,\ k = 1, 2, \ldots, n).$$

(ii) The estimation of the probability function of K_n in case the parameter θ is unknown:
The necessary values for the approximation II are

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

Since we consider the case the parameter θ is unknown, we take MLE $\hat{\theta}$ as the estimator θ. Then, by (9) and the above relations, we have

$$\lambda_{n-1} = k - 1, \quad \lambda_{2,n-1}^{**} := \lambda_{2,n-1} = \hat{\theta}^2[\psi'(\hat{\theta} + 1) - \psi'(\hat{\theta} + n)].$$

Putting

$$(n-1)^{**} = \lfloor (k-1)^2 / \lambda_{2,n-1}^{**} \rfloor,\ p^{**} = (k-1)/(n-1)^{**},$$

we obtain the estimator of probability function of K_n given by

$$g_B(x - 1; (n-1)^{**}, p^{**}) \ (x = 1, 2, \ldots, (n-1)^{**} + 1).$$

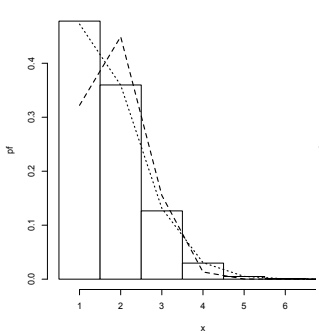

Fig. 5: $n = 250, \theta = 0.125$ N(dash) and IV.1(dot)

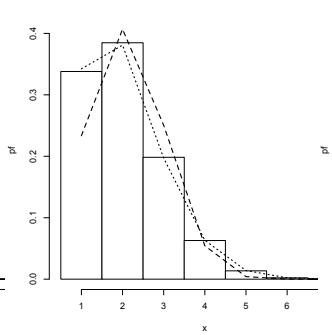

Fig. 6: $n = 50, \theta = 0.25$ N(dash) and IV.1(dot)

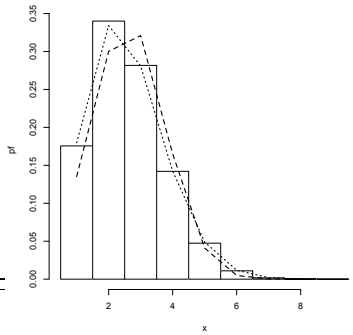

Fig. 7: $n = 25, \theta = 0.5$ N(dash) and IV.1(dot)

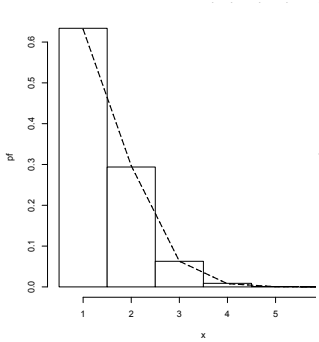

Fig. 8: $n = 25, \theta = 0.125$ sPo(dash) and II(dot)

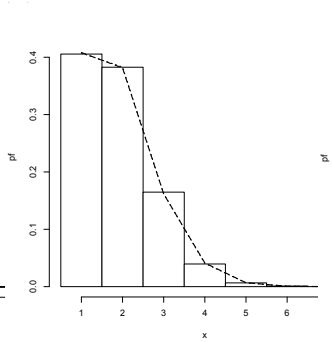

Fig. 9: $n = 25, \theta = 0.25$ sPo(dash) and II(dot)

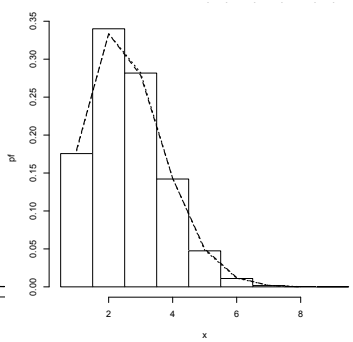

Fig. 10: $n = 25, \theta = 0.5$ sPo(dash) and II(dot)

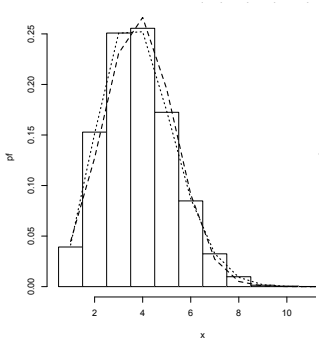

Fig. 11: $n = 25, \theta = 1$ N(dash) and IV.1(dot)

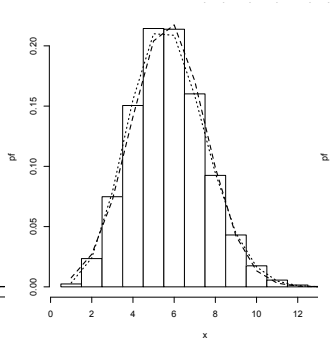

Fig. 12: $n = 25, \theta = 2$ N(dash) and IV.1(dot)

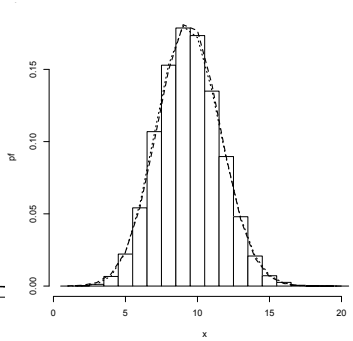

Fig. 13: $n = 25, \theta = 5$ N(dash) and IV.1(dot)

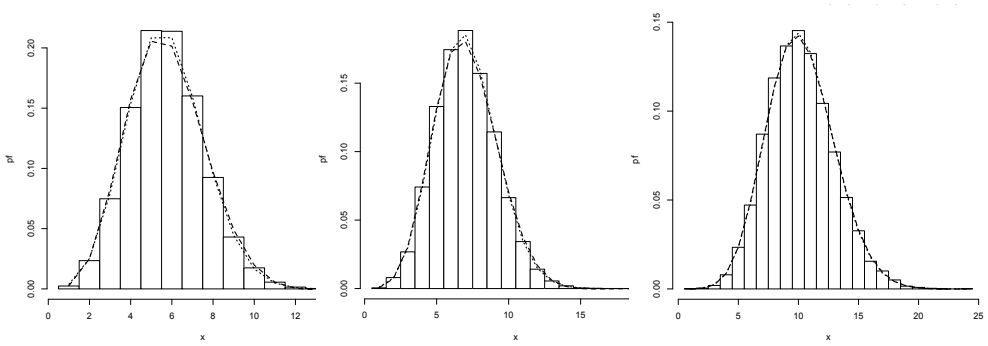

Fig. 14: $n = 25, \theta = 2$ sPo(dash) and II(dot)

Fig. 15: $n = 50, \theta = 2$ sPo(dash) and II(dot)

Fig. 16: $n = 250, \theta = 2$ sPo(dash) and II(dot)

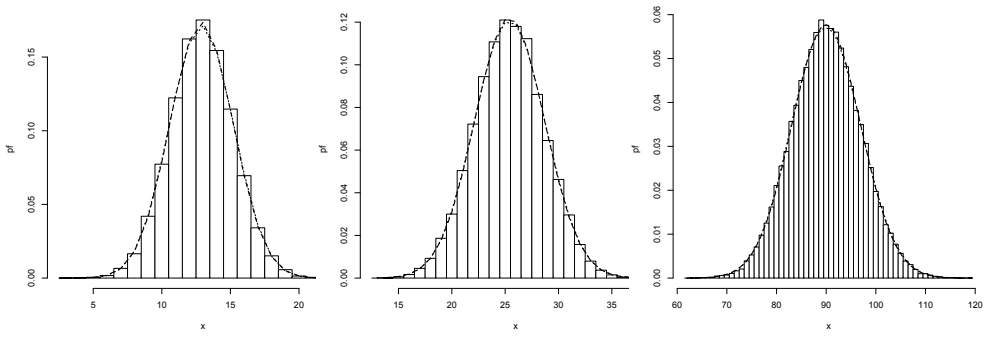

Fig. 17: $n = 25, \theta = 10$ N(dash) and IV.1(dot)

Fig. 18: $n = 50, \theta = 20$ N(dash) and IV.1(dot)

Fig. 19: $n = 250, \theta = 50$ N(dash) and IV.1(dot)

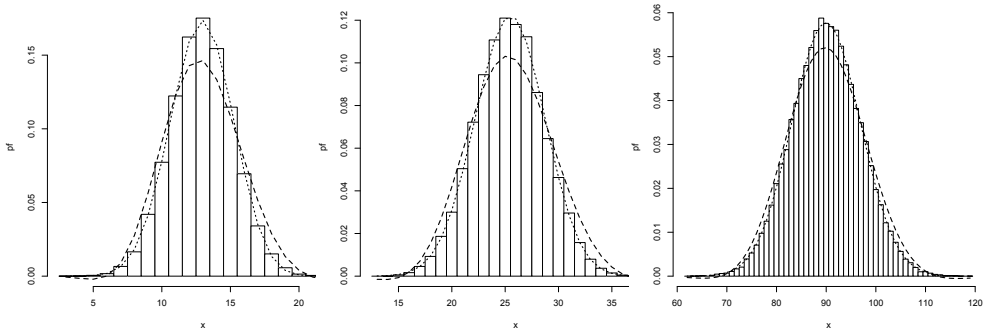

Fig. 20: $n = 25, \theta = 10$ sPo(dash) and II(dot)

Fig. 21: $n = 50, \theta = 20$ sPo(dash) and II(dot)

Fig. 22: $n = 250, \theta = 50$ sPo(dash) and II(dot)

Acknowledgement

The author expresses gratitude to an anonymous referee for his careful reading of the manuscript and valuable comments. 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.

Crane, H. (2016). The ubiquitous Ewens sampling formula. *Statististical Science*, **31**, 1–19.

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. (2001). Binomial approximation to the Poisson binomial distribution: The Krawtchouk expansion. *Theory of Probability and its Applications*, **45**, 258–272.

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

Yamato, H. (2017a). Shifted Binomial approximation for the Ewens sampling formula. *Bulletin of Informatics and Cybernetics*, **49**, 81–88.

Yamato, H. (2017b). Poisson approximations for sum of bernoulli random variables and its application to Ewens sampling formula. *Journal of the Japan Statistical Society*, **47**, 2, 187–195.

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 April 25, 2018
Revised August 30, 2018