POISSON APPROXIMATIONS FOR SUM OF BERNOULLI RANDOM VARIABLES AND ITS APPLICATION TO EWENS SAMPLING FORMULA

Hajime Yamato*

The Ewens sampling formula is well-known as a distribution of a random partition of the set of integers $\{1, 2, \dots, n\}$. We give the condition that the number K_n of distinct components of the formula converges to the shifted Poisson distribution. Based on this convergence, we give the new approximations to the distribution of K_n , which are different from the approximations by Arratia *et al.* (2000, 2003). The formers are better than the latters. This is shown by comparing the bounds for the total variation distances between the distributions of the approximations and the distribution of K_n . Several examples are given to illustrate the results.

Key words and phrases: Bernoulli random variable, Charlier polynomial, Ewens sampling formula, Poisson distribution, shifted distribution.

1. Introduction

Ewens (1972) discovered a distribution of a random partition of the set of integers $\{1, 2, \dots, n\}$, 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. The number K_n of distinct components of the Ewens sampling formula has the distribution whose probability function given by $P(K_n = k) = |s(n, k)|\theta^k/\theta^{[n]}$ ($k = 1, 2, \dots, n$), where $0 < \theta < \infty$, $\theta^{[n]} = \theta(\theta + 1) \cdots (\theta + n - 1)$ and $|s(n, k)|$ 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 are written using the digamma function and trigamma function, and these function are included in the programming language R, the normal approximation to the distribution $\mathcal{L}(K_n)$ of K_n is easily obtained using R (Yamato *et al.* (2015)).

On the other hand, the Poisson approximation to $\mathcal{L}(K_n)$ are studied by Arratia *et al.* (2000) in detail with respect to Logarithmic combinatorial structure including Ewens sampling formula. By the method different from Arratia *et al.* (2000), we approach to the problem of Poisson approximation to $\mathcal{L}(K_n)$.

We approach to the problem by using the sum of Bernoulli random variables. In the Section 2, we quote the previous works for the sum of independent but non-identically distributed Bernoulli random variables. We give the elementally

proof for the convergence in distribution to the Poisson distribution. In addition, we quote the approximation by using the Charlier polynomials.

In the Section 3, we treat the Ewens sampling formula. We give the necessary and sufficient condition that K_n converges to the shifted Poisson distribution. In addition, we propose new approximations to $\mathcal{L}(K_n)$, one of which uses the Charlier polynomial. They are better than the approximations of Arratia *et al.* (2000, 2003). This is shown by comparing the bounds for the total variation distances between the distributions of the approximations and $\mathcal{L}(K_n)$. These results are illustrate by several examples.

2. Sum of independent Bernoulli random variables

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

$$S_n = \sum_{j=1}^n X_j, \quad \lambda_n = \sum_{j=1}^n p_j \quad \text{and} \quad \lambda_{2,n} = \sum_{j=1}^n p_j^2.$$

The total variation distance d_{TV} on probabilities Q_1 and Q_2 over $\{0, 1, 2, \dots\}$ is defined by

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

The distribution $\mathcal{L}(S_n)$ of S_n is approximated by the Poisson distribution $Po(\lambda_n)$ with parameter $\lambda_n (> 0)$. For the total variation distance between $\mathcal{L}(S_n)$ and $Po(\lambda_n)$, Le Cam (1960; Theorem 1) gives the upper bound as follows:

$$(2.1) \quad d_{TV}(\mathcal{L}(S_n), Po(\lambda_n)) \leq \lambda_{2,n}.$$

Barbour and Hall (1984; Theorem 1) improves this as follows:

$$(2.2) \quad d_{TV}(\mathcal{L}(S_n), Po(\lambda_n)) \leq \lambda_n^{-1}(1 - e^{-\lambda_n})\lambda_{2,n}$$

and gives the lower bound (Theorem 2),

$$(2.3) \quad d_{TV}(\mathcal{L}(S_n), Po(\lambda_n)) \geq \frac{1}{32} \min\{1, \lambda_n^{-1}\}\lambda_{2,n}.$$

For the convergence in distribution of S_n , using the generating function Feller (1968; p. 282) shows that under the condition that $\max_{1 \leq j \leq n} p_j \rightarrow 0$ and the sum $\lambda_n = \lambda$ remains constant, S_n converge in distribution to $Po(\lambda)$. Hereafter this convergence is abbreviated to $S_n \xrightarrow{d} Po(\lambda)$. Under the condition that $\max_{1 \leq j \leq n} p_j \rightarrow 0$ and $\lambda_n \rightarrow \lambda$, Billingsley (1995; Theorem 23.2) shows $S_n \xrightarrow{d} Po(\lambda)$ by the point-wise convergence of the p.f. (probability function) of S_n to the p.f. of $Po(\lambda)$.

The necessary and sufficient condition for the convergence in distribution of S_n to $Po(\lambda)$ is given by Wang (1993; Theorem 3).

PROPOSITION 2.1 (Wang (1993)). *For the sum S_n of independent Bernoulli random variables, $S_n \xrightarrow{d} Po(\lambda)$ ($0 < \lambda < \infty$) if and only if*

$$(2.4) \quad \lambda_n \rightarrow \lambda \quad \text{and} \quad \lambda_{2,n} \rightarrow 0 \quad \text{as} \quad n \rightarrow \infty.$$

While Wang (1993) proves the sufficient condition by the point-wise convergence of the p.f., it is easily proved by the method of moments as follows:

PROOF. We first suppose (2.4). The factorial moment generating function of S_n is $F(t) = E[(1+t)^{S_n}] = \prod_{j=1}^n (1+p_j t)$ 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

$$E[S_n^{(m)}] = \sum_{1 \leq j_1 \neq \cdots \neq j_m \leq n} p_{j_1} \cdots p_{j_m} \quad (m = 1, 2, \dots, n),$$

where the summation of the right-hand side is taken over all positive integers $j_1, \dots, j_m$ satisfying $1 \leq j_1 \neq \cdots \neq j_m \leq n$. We note

$$(2.5) \quad E[S_n^{(m)}] = \left(\sum_{j=1}^n p_j \right)^m - I_{n,m}, \quad I_{n,m} = \sum_{l_1 + \cdots + l_n = m}^* p_1^{l_1} \cdots p_n^{l_n},$$

where the last summation is taken over all non-negative integers $l_1, \dots, l_n$ satisfying $l_1 + \cdots + l_n = m$, except for the case that m terms among $l_1, \dots, l_n$ are equal to 1 and the other $n-m$ terms are equal to 0. Hence, for $I_{n,m}$, at least one of $l_1, \dots, l_n$ is equal to 2. Among the sum $I_{n,m}$, for example, we consider the following sum I_1^* such that the only one term is equal to 2 and all the other non-zero $m-2$ terms are equal to 1. Then we have

$$I_1^* = \sum_{1 \leq j_1 \neq \cdots \neq j_{m-1} \leq n} p_{j_1}^2 p_{j_2} \cdots p_{j_{m-1}} \leq \left(\sum_{j=1}^n p_j^2 \right) \left(\sum_{j=1}^n p_j \right)^{m-2}.$$

Therefore, under the condition (2.4), I_1^* tends to zero as $n \rightarrow \infty$. Similarly, we can show the other sums of $I_{n,m}$ tends to zero and hence $I_{n,m}$ tends to zero, as $n \rightarrow \infty$. Thus, by (2.4) and (2.5), for $m = 1, 2, \dots$ we get

$$E[S_n^{(m)}] \rightarrow \lambda^m \quad \text{as} \quad n \rightarrow \infty,$$

so that $S_n \xrightarrow{d} Po(\lambda)$.

On the contrary, we suppose $S_n \xrightarrow{d} Po(\lambda)$. Then both the mean $E[S_n] = \lambda_n$ and the variance $V[S_n] = \lambda_n - \lambda_{2,n}$ converge to λ as $n \rightarrow \infty$. Therefore we have (2.4). $\square$

Remark 2.1. The first part of the above proof is more easily shown using the cumulants. The cumulants κ_m ($m = 1, 2, \dots$) of S_n are given by

$$\kappa_1 = \sum_{i=1}^n p_i, \quad \kappa_2 = \sum_{i=1}^n p_i - \sum_{i=1}^n p_i^2, \quad \kappa_3 = \sum_{i=1}^n p_i - 3 \sum_{i=1}^n p_i^2 + 2 \sum_{i=1}^n p_i^3,$$

and

$$\kappa_m = \sum_{i=1}^n p_i + c_{m,2} \sum_{i=1}^n p_i^2 + \dots + c_{m,m} \sum_{i=1}^n p_i^m \quad (m = 4, 5, \dots),$$

where $c_{m,2}, \dots, c_{m,m}$ are constants not depending on n . By $\lambda_{2,n} = \sum_{i=1}^n p_i^2 \rightarrow 0$, we have $\sum_{i=1}^n p_i^m \rightarrow 0$ ($m = 3, 4, \dots$), as $n \rightarrow \infty$. Hence, for $m = 1, 2, \dots$, κ_m converges to λ as $n \rightarrow \infty$, so that $S_n \xrightarrow{d} Po(\lambda)$.

Remark 2.2. Next, we shall examine Proposition 2.1 by the total variation distances. By the relations (5), (8) and (9) of Yannaros (1991), we have

$$1 - \rho \leq d_{TV}(Po(\lambda_n), Po(\lambda)) \leq \sqrt{1 - \rho^2}, \quad \rho = \exp \left\{ -\frac{1}{2}(\sqrt{\lambda_n} - \sqrt{\lambda}) \right\}.$$

Therefore, the condition $\lambda_n \rightarrow \lambda$ of (2.4) is equivalent to $d_{TV}(Po(\lambda_n), Po(\lambda)) \rightarrow 0$, as $n \rightarrow \infty$. On the other hand, by (2.2) and (2.3), $\lambda_{2,n} \rightarrow 0$ of (2.4) is equivalent to $d_{TV}(Po(S_n), Po(\lambda_n)) \rightarrow 0$, as $n \rightarrow \infty$. Applying these relations to the right-hand side of the inequality,

$$d_{TV}(Po(S_n), Po(\lambda)) \leq d_{TV}(Po(S_n), Po(\lambda_n)) + d_{TV}(Po(\lambda_n), Po(\lambda)),$$

under the condition (2.4) we have $d_{TV}(Po(S_n), Po(\lambda)) \rightarrow 0$ ($n \rightarrow \infty$), and hence $S_n \xrightarrow{d} Po(\lambda)$.

While as the approximation to $\mathcal{L}(S_n)$ we can use $Po(\lambda_n)$ by Proposition 2.1, we shall consider the more accurate approximation. Let $p(x, \lambda) = e^{-\lambda} \lambda^x / x!$ be the p.f. of Poisson distribution with parameter λ . The j th Charlier polynomial $C_j(x, \lambda)$ is defined by

$$C_j(x, \lambda) = \frac{d^j}{d\lambda^j} p(x, \lambda) / p(x, \lambda).$$

The Charlier polynomials are orthogonal with respect to the p.f. $p(x, \lambda)$. That is,

$$\sum_{x=0}^{\infty} C_k(x, \lambda) C_l(x, \lambda) p(x, \lambda) = \frac{k!}{\lambda^k} \delta_{k,l},$$

where $\delta_{k,l} = 1$ if $k = l$ and $= 0$ if $k \neq l$. The j th Charlier polynomial can be expressed as

$$C_j(x, \lambda) = \sum_{i=0}^j \binom{j}{i} (-1)^{j-i} \frac{x(x-1) \cdots (x-i+1)}{\lambda^i}.$$

Particularly, we have

$$C_1(x, \lambda) = \frac{x - \lambda}{\lambda} \quad \text{and} \quad C_2(x, \lambda) = \frac{x^2 - (2\lambda + 1)x + \lambda^2}{\lambda^2}.$$

For the Charlier polynomials, see, for example, Barbour *et al.* (1992; p. 175), Takeuchi (1975), and Zacharovas and Hwang (2010). We put

$$(2.6) \quad \alpha(m, n, \lambda) = p(m, \lambda_n) \left(1 - \frac{\lambda_{2,n}}{2} C_2(m, \lambda_n) \right).$$

Zacharovas and Hwang (2010; Theorem 4.2) gives

$$(2.7) \quad \frac{1}{2} \sum_{m=0}^{\infty} |P(S_n = m) - \alpha(m, n, \lambda)| \leq \frac{c_1 \eta_n^2}{(1 - \eta_n)^{5/2}} + \frac{c_2 \lambda_{3,n}}{\lambda_n^{3/2} (1 - \eta_n)^2},$$

where $\eta_n = \lambda_{2,n}/\lambda_n < 1$, $\lambda_{3,n} = \sum_{j=1}^n p_j^3$ and c_1, c_2 are positive constants. As the approximation to $\mathcal{L}(S_n)$, we can use $\alpha(m, n, \lambda)$. This approximation $\alpha(m, n, \lambda)$ is also obtained from Takeuchi (1975; Theorem 1.6 and Example 1.6 with c_1, c_2), where a_i in Takeuchi (1975) should be changed to c_i for $i = 1, 2, 3, 4$.

3. Ewens sampling formula

For the logarithmic combinatorial structure including the Ewens sampling formula, the Poisson approximation to the distribution of the number K_n of distinct components are derived by Arratia *et al.* (2000; Theorem 5.4, Corollary 5.5 and Remark, 2003; Theorem 8.15) in detail. Here, we quote their approximations in the case of the Ewens sampling formula. Following their notations, let $\tau_n = \theta[\psi(n+1) - \psi(\theta+1)]$, $a_n = -\theta^2 \psi'(\theta+1)$ and

$$\nu_n\{s+1\} = p(s, \tau_n) \left(1 + \frac{a_n}{2\tau_n^2} \{(s - \tau_n)^2 - \tau_n\} \right),$$

where $\psi(t) = \Gamma'(t)/\Gamma(t)$ is the digamma function and $\psi'(t)$ is the trigamma function. Since a_n is not integer in general, we quote the followings:

$$(3.1) \quad \nu_n, \quad Po(1 + \tau_n) \quad \text{and} \quad 1 + Po(\tau_n).$$

We shall derive the approximation different from (3.1), using the results of Section 2. For the Ewens sampling formula, the number K_n of distinct components can be expressed as

$$K_n = \xi_1 + \xi_2 + \cdots + \xi_n \quad (n = 1, 2, \dots),$$

where $\xi_1, \xi_2, \dots, \xi_n$ are independent, and

$$P(\xi_j = 1) = \frac{\theta}{\theta + j - 1}, \quad P(\xi_j = 0) = \frac{j - 1}{\theta + j - 1} \quad (j = 1, 2, \dots; 0 < \theta < \infty).$$

Since $\xi_1 = 1$ a.s. (almost surely), K_n can be expressed as

$$(3.2) \quad K_n = 1 + L_n \quad \text{a.s.} \quad (n = 2, 3, \dots),$$

where $L_n = \xi_2 + \dots + \xi_n$. We let

$$(3.3) \quad \mu_n = \sum_{i=2}^n \frac{\theta}{\theta + i - 1} = \theta[\psi(\theta + n) - \psi(\theta + 1)],$$

and

$$(3.4) \quad \mu_{2,n} = \sum_{i=2}^n \left(\frac{\theta}{\theta + i - 1} \right)^2 = \theta^2[\psi'(\theta + 1) - \psi'(\theta + n)].$$

PROPOSITION 3.1. *For the number K_n of distinct components of the Ewens sampling formula, $K_n \xrightarrow{d} 1 + Po(\lambda)$ ($0 < \lambda < \infty$) if and only if*

$$(3.5) \quad \theta \log n \rightarrow \lambda \quad \text{as } n \rightarrow \infty \text{ and } \theta \rightarrow 0.$$

PROOF. By the inequalities $\sum_{i=2}^n 1/(\theta + i) < \int_1^n 1/(\theta + x)dx < \sum_{i=2}^n 1/(\theta + i - 1)$, we have

$$(3.6) \quad \sum_{i=2}^n \frac{\theta}{\theta + i} < \theta[\log(\theta + n) - \log(\theta + 1)] < \sum_{i=2}^n \frac{\theta}{\theta + i - 1}.$$

Similarly, we have

$$(3.7) \quad \mu_n = \sum_{i=2}^n \frac{\theta}{\theta + i - 1} < \theta[\log(\theta + n - 1) - \log \theta].$$

By (3.6) and (3.7), $\mu_n \rightarrow \lambda$ as $n \rightarrow \infty$ is equivalent to (3.5). Since $\mu_{2,n} < \theta^2 \pi^2/6$, under (3.5) we have $\mu_{2,n} \rightarrow 0$. Therefore, by Proposition 2.1 with μ_n and $\mu_{2,n}$ instead of λ_n and $\lambda_{2,n}$, respectively, we get that $L_n \xrightarrow{d} Po(\lambda)$ if and only if (3.5). Thus, by (3.2) we have Proposition 3.1. $\square$

By Proposition 2.1, (2.6) and Proposition 3.1, as the approximation to the distribution of K_n associated with the Ewens sampling formula we take

$$(3.8) \quad 1 + Po(\mu_n) \quad \text{and} \quad 1 + p(m, \mu_n) \left(1 - \frac{\mu_{2,n}}{2} C_2(m, \mu_n) \right).$$

We note $\sum_{i=2}^n (\theta + i - 1)^{-3} < \sum_{i=2}^n (\theta + i - 1)^{-2} < \pi^2/6$ and $\sum_{i=2}^n (\theta + i - 1)^{-1} \sim \log n$. Since $\theta \sim 1/\log n$ under the condition (3.5), by (2.7) the total variation distances between $\mathcal{L}(K_n)$ and (3.8) are

$$(3.9) \quad d_{TV}(\mathcal{L}(K_n), 1 + Po(\mu_n)) = d_{TV}(\mathcal{L}(L_n), Po(\mu_n)) = O(\log^{-2} n),$$

and

$$(3.10) \quad \frac{1}{2} \sum_{m=0}^{\infty} \left| P(L_n = m) - p(m, \mu_n) \left(1 - \frac{\mu_{2,n}}{2} C_2(m, \mu_n) \right) \right| = O(\log^{-3} n),$$

respectively. The total variation distances between $\mathcal{L}(K_n)$ and (3.1) are

$$(3.11) \quad d_{TV}(\mathcal{L}(K_n), \nu_n) = O(\log^{-3/2} n),$$

$$(3.12) \quad d_{TV}(\mathcal{L}(K_n), Po(1 + \tau_n)) = O(\log^{-1} n),$$

$$(3.13) \quad d_{TV}(\mathcal{L}(K_n), 1 + Po(\tau_n)) = O(\log^{-3/2} n) + |a_n|/\sqrt{2\pi e},$$

respectively (Arratia *et al.* (2000; Theorem 5.4 and its Corollary and Remark)).

With respect to the order of the total variation distance, (3.9) and (3.10) are smaller than (3.11), (3.12) and (3.13). Thus, the approximations (3.8) are better than (3.1). Probably, this is caused by the fact that the approximations (3.1) distribute around $1 + \tau_n = 1 + \theta[\psi(n+1) - \psi(\theta+1)]$, while the approximations (3.8) distribute around $1 + \mu_n = 1 + \theta[\psi(n+\theta) - \psi(\theta+1)]$ which is equal to the mean $E(K_n)$. For a small θ , τ_n is almost equal to μ_n , depending on n . As θ increases, τ_n becomes smaller than μ_n , depending on n . We illustrate this by Figs. 1 and 2 for $n = 25, 50$. Thus, for a small θ , two approximations (3.1) and (3.8) are not so much different. These are illustrated by Figs. 3 and 4, for $\theta = 0.5$ and $n = 25, 50$. For a large θ , the approximations (3.1) have more weight on the left hand side of $\mathcal{L}(K_n)$. These are illustrated by Figs. 5, 6, 7 and 8, for $\theta = 5, n = 25$ and $\theta = 10, n = 50$. It can be seen that as the Poisson approximation to $\mathcal{L}(K_n)$, (3.8) is better than (3.1).

In the following figures, we draw the p.f. of K_n by simulation using R, as bar graph. For the approximation (3.1), we draw ν_n by solid line, $Po(1 + \tau_n)$ by dashed line and $1 + Po(\tau_n)$ by dotted line, in Figs. 3, 5 and 7. For the approximation (3.8), we draw $1 + Po(\mu_n)$ by dotted line and $1 + p(m; \mu_n)(1 - \mu_{2,n}C_2(m, \mu_n)/2)$ by solid line, in Figs. 4, 6 and 8. In Figs. 5, 6, 7 and 8, the tails of ν_n and $1 + p(m; \mu_n)(1 - \mu_{2,n}C_2(m, \mu_n)/2)$ are omitted because of their negative values.

Figure 1. τ_n and μ_n for $n = 25$.

Figure 2. τ_n and μ_n for $n = 50$.

Figure 3. (3.1) with $n = 25$, $\theta = 0.5$. Figure 4. (3.8) with $n = 25$, $\theta = 0.5$. Figure 5. (3.1) with $n = 25$, $\theta = 5$. Figure 6. (3.8) with $n = 25$, $\theta = 5$. Figure 7. (3.1) with $n = 50$, $\theta = 10$. Figure 8. (3.8) with $n = 50$, $\theta = 10$.

Acknowledgements

The author expresses gratitude to two anonymous referees for their 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, *Ann. Appl. Probab.*, **10**, 331–361.
- Arratia, R., Barbour, A. D. and Tavaré, S. (2003). *Logarithmic Combinatorial structure: A Probabilistic Approach*, European Mathematical Society.
- Barbour, A. D. and Hall, P. (1984). On the rate of Poisson convergence, *Math. Proc. Cambridge Philos. Soc.*, **95**, 473–480.
- Barbour, A. D., Holst, L. and Janson, S. (1992). *Poisson Approximation*, Clarendon Press, Oxford.
- Billingsley, P. (1995). *Probability and Measure*, 3rd ed., John Wiley & Sons, New York.
- Ewens, W. J. (1972). The sampling theory of selectively neutral alleles, *Theor. Popul. Biol.*, **3**, 87–112.
- Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, Vol. I, 3rd ed., John Wiley & Sons, New York.
- 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.
- Le Cam, L. (1960). An approximation theorem for the Poisson binomial distribution, *Pacific J. Math.*, **10**, 1181–1197.
- Takeuchi, K. (1975). *Approximation of Probability Distributions* (in Japanese), Kyouiku Shuppan, Tokyo.
- Wang, Y. H. (1993). On the number of successes in independent trials, *Statistica Sinica*, **3**, 295–312.
- Yamato, H., Nomachi, T. and Toda, K. (2015). Approximate distributions for the number of distinct components of the Ewens sampling formula and its applications, *Bull. Inform. Cybern.*, **47**, 69–81.
- Yannaros, N. (1991). Poisson approximation for random sums of Bernoulli random variables, *Stat. Probab. Lett.*, **11**, 161–165.
- Zacharovas, V. and Hwang, H.-K. (2010). A Charlier-Parseval approach to Poisson approximation and its applications, *Lithuania Mathematical J.*, **50**, 88–119.