# ASYMPTOTIC DISTRIBUTION OF NUMBER OF DISTINCT OBSERVATIONS AMONG A SAMPLE FROM MIXTURE OF DIRICHLET PROCESSES

Yamato, Hajime
Kagoshima University : Professor Emeritus

https://doi.org/10.5109/1495410

出版情報：Bulletin of informatics and cybernetics. 44, pp.41-47, 2012-12. Research Association of Statistical Sciences

# ASYMPTOTIC DISTRIBUTION OF NUMBER OF DISTINCT OBSERVATIONS AMONG A SAMPLE FROM MIXTURE OF DIRICHLET PROCESSES

**by**

**Hajime Yamato**

---

*Reprinted from the Bulletin of Informatics and Cybernetics Research Association of Statistical Sciences, Vol.44*

FUKUOKA, JAPAN
2012

*Bulletin of Informatics and Cybernetics, Vol. 44, 2012*

# ASYMPTOTIC DISTRIBUTION OF NUMBER OF DISTINCT OBSERVATIONS AMONG A SAMPLE FROM MIXTURE OF DIRICHLET PROCESSES

By

**Hajime Yamato***

**Abstract**

Let $X_1, X_2, \cdots, X_n$ be a sample of size $n$ from a random discrete distribution $\mathcal{P}$ on the real line $\mathbb{R}$. If we consider $i$ and $j$ are equivalent in case of $X_i = X_j$, this equivalence relation give a random partition of $\mathbb{N}_n = \{1, 2, \cdots, n\}$. In the case where $\mathcal{P}$ is given by a mixture of Dirichlet processes, we discuss the convergence in distribution of the number $K_n$ of distinct components of the random partition of $\mathbb{N}_n$.

*Key Words and Phrases:* Mixture of Dirichlet processes, Normal approximation, Poisson distribution, random partition, smoothing lemma.

## 1. Introduction

Let $G_0$ be a continuous distribution on the real line $\mathbb{R}$ and $\theta$ be a positive constant. Let $\mathcal{B}$ be the $\sigma$-field which consists of the subsets of $\mathbb{R}$. Let the random distribution $\mathcal{P}$ have the Dirichlet process $\mathcal{D}(\theta G_0)$ on $(\mathbb{R}, \mathcal{B})$ with parameter $\theta G_0$ . Let $V_j$ $(j = 1, 2, \ldots)$ be a sequence of independent and identically distributed (i.i.d.) random variables with the distribution $G_0$, and $W_j (j = 1, 2, \ldots)$ be a sequence of i.i.d. random variables with the beta distribution $Be(1, \theta)$. We assume that $V_j$ $(j = 1, 2, \ldots)$ and $W_j (j = 1, 2, \ldots)$ are independent. We put $p_1 = W_1$ and $p_j = W_j (1 - W_1) \cdots (1 - W_{j-1})$ $(j = 2, 3, \cdots)$. Then, we can write $\mathcal{P}(B) = \sum_{j=1}^{\infty} p_j \delta_{V_j}(B)$ for any $B \in \mathcal{B}$, where $\delta_V(B) = 1$ if $V \in B$ and 0 otherwise (Sethuraman (1994)). Thus $\mathcal{P}$ $(\in \mathcal{D}(\theta G_0))$ is discrete almost surely (a.s.). A sample of size $n$ from $\mathcal{P}$ gives the random partition of $\mathbb{N}_n = \{1, 2, \cdots, n\}$, whose distribution does not depend on $V_j$ $(j = 1, 2, \ldots)$ given $\mathcal{P}$. Thus the distribution depends on $\theta$ and does not depend on $G_0$. The distribution is well-known as Ewens sampling formula or Multivariate Ewens distribution. Let $K_n$ be the number of distinct components of the random partition. The distribution of $K_n$ is given by

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

where $\left[ \begin{array}{c} n \\ k \end{array} \right]$ is a Stirling number of the third kind (or an absolute Stirling number of the first kind) and $\theta^{[n]} = \theta(\theta + 1) \cdots (\theta + n - 1)$. (See, for example, Antoniak (1974) and Johnson et al. (1997).)

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

Hereafter we consider $\theta$ as a positive random variable having a distribution $\gamma$. Given $\theta$, let the random discrete (a.s.) distribution $\mathcal{P}$ have the Dirichlet process $\mathcal{D}(\theta G_0)$ on ($\mathbb{R}$, $\mathcal{B}$) with parameter $\theta G_0$. Then this random discrete (a.s.) distribution $\mathcal{P}$ has a mixture of Dirichlet processes $\mathcal{D}(\theta G_0)$ with the mixing distribution $\gamma$ (Antoniak (1974)). The number $K_n$ of distinct components of the random partition based on a sample of size $n$ associated with this mixture of Dirichlet processes, have the distribution (1), given $\theta$. Thus, for the mixture of Dirichlet processes $\mathcal{D}(\theta G_0)$ with the mixing distribution $\gamma$, $K_n$ has the distribution

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

where $E_\gamma(\cdot)$ denotes the expectation with respect to the distribution $\gamma$ of the random variable $\theta$. We note that the relation between (1) and (2) corresponds to (ii) of Theorem 12 which is the characterization of the random partition which are consistent and exchangeable (Gnedin and Pitman (2006)).

The purpose of this paper is to show that $K_n/\log n$ converges in distribution to $\gamma$ as $n \to \infty$ and its order is $O(\log^{-1/3} n)$, for the mixture of Dirichlet processes $\mathcal{D}(\theta G_0)$ with the mixing distribution $\gamma$.

## 2. Convergence in distribution of $K_n$

The total variation distance between the distribution $\mathcal{L}(X)$ and $\mathcal{L}(Y)$ of discrete nonnegative random variables $X$ and $Y$, $\| \mathcal{L}(X) - \mathcal{L}(Y) \|$, is defined by

$$\| \mathcal{L}(X) - \mathcal{L}(Y) \| = \sup_{B \subset \mathbb{Z}_+} \left| P(X \in B) - P(Y \in B) \right|$$

where $\mathbb{Z}_+ = \{0, 1, 2, \cdots\}$. For each $n = 1, 2, \cdots$, we put

$$H_{\theta n} = \sum_{j=1}^{n} \frac{1}{\theta + j - 1}.$$

Let the random variables $P_n$ and $P_n^*$ have the Poisson distributions $\mathrm{Po}(\theta H_{\theta n})$ and $\mathrm{Po}(\theta \log n)$, respectively, given $\theta$. In order to discuss the convergence in distribution of $K_n/\log n$, at first we see the total variation distances $\| \mathcal{L}(K_n) - \mathcal{L}(P_n) \|$ and $\| \mathcal{L}(P_n) - \mathcal{L}(P_n^*) \|$. Thus we see the total variation distance $\| \mathcal{L}(K_n) - \mathcal{L}(P_n^*) \|$ and Lemma 2.2. These are shown in the subsection 2.1. Then in the subsection 2.2 we show that $P_n^*/\log n$ converges in distribution to $\gamma$ and have Lemma 2.3. By Lemmas 2.2 and 2.3, we have the following.

Proposition 2.1. *Let $K_n$ be the number of distinct observations among a sample of size $n$ associated with the mixture of Dirichlet processes $\mathcal{D}(\theta G_0)$ with the mixing distribution $\gamma$, where $\gamma$ is a distribution of the positive random variable $\theta$ and $G_0$ is a continuous distribution on $\mathbb{R}$. We suppose that the probability density function (p.d.f.) of $\gamma$ is bounded, and that $E_\gamma(\theta^{-1})$ and $E_\gamma(\theta^2)$ exist. Then we have*

$$\sup_{-\infty < x < \infty} \left| P\left( \frac{K_n}{\log n} \le x \right) - \gamma(x) \right| = O\left( \frac{1}{\sqrt[3]{\log n}} \right). \tag{3}$$

## 2.1. $K_n$ and Poisson distribution

### 2.1.1. $K_n$ and $P_n$

Given $\theta$, let random variables $\xi_j$ $(j = 1, 2, \cdots)$ be independent and take the value 0,1 with the probabilities given by

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

Then, given $\theta$, that is, for Ewens sampling formula, $K_n$ can be written as

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

(see, for example, Johnson et al. (1997)). Given $\theta$, for the total variation distance between $\mathcal{L}(K_n|\theta)$ and $\mathcal{L}(P_n|\theta)$, we have

$$\| \mathcal{L}(K_n|\theta) - \mathcal{L}(P_n|\theta) \| \leq \lambda^{-1}(1 - e^{-\lambda}) \sum_{j=1}^{n} p_j^2 \tag{4}$$

where $\mathcal{L}(X|\theta)$ denotes the conditional distribution of $X$ given $\theta$, and

$$p_j = \frac{\theta}{\theta + j - 1}, \quad \lambda = \sum_{j=1}^{n} p_j = \theta H_{\theta n}.$$

(With respect to the inequality (4) for the total variation distance between the distribution of sum of Bernoulli random variables and the Poisson distribution, see Barbour and Hall (1984), Theorem 2.)

Since $\lambda > 0$, we have

$$0 < 1 - e^{-\lambda} < 1.$$

We also have

$$\sum_{j=1}^{n} p_j^2 \leq 1 + \theta^2 \sum_{j=1}^{n-1} \frac{1}{j^2} \leq 1 + \frac{\pi^2}{6}\theta^2.$$

We note that

$$H_{\theta n} > \int_0^n \frac{1}{\theta + x} dx > \log n \ \text{ for } \ 0 < \theta < 1.$$

For $\theta \geq 1$, since $\theta/(\theta + j - 1) \geq 1/j$ $(j = 1, 2, \cdots)$, we have $\theta H_{\theta n} \geq H_n$, where $H_n = \sum_{j=1}^{n}(1/j)$ is the harmonic number. Since $H_n > \log n$, we have $\theta H_{\theta n} > \log n$ for $\theta \geq 1$. Thus we have

$$\lambda^{-1} = \frac{1}{\theta H_{\theta n}} \leq \left\{ \begin{array}{ll} 1/(\theta \log n) & (0 < \theta < 1) \\ 1/\log n & (\theta \geq 1) \end{array} \right. .$$

Hence, we have

$$\| \mathcal{L}(K_n|\theta) - \mathcal{L}(P_n|\theta) \| \leq \frac{c(\theta)}{\log n}$$

where

$$c(\theta) = \frac{1}{\theta} + \frac{\pi^2}{6}\theta \ \ (0 < \theta < 1), \quad = 1 + \frac{\pi^2}{6}\theta^2 \ \ (\theta \geq 1).$$

Thus, if $E_\gamma(\theta^{-1})$ and $E_\gamma(\theta^2)$ exist, then we have

$$\| \mathcal{L}(K_n) - \mathcal{L}(P_n) \| \leq E_\gamma \| \mathcal{L}(K_n|\theta) - \mathcal{L}(P_n|\theta) \| = O\left(\frac{1}{\log n}\right). \tag{5}$$

**2.1.2. $P_n$ and $P_n^*$**

We consider the total variation distance between the Poisson distribution $Po(\theta H_{\theta n})$ and $Po(\theta \log n)$, given $\theta$. Since, given $\theta$, $P_n$ and $P_n^*$ have $Po(\theta H_{\theta n})$ and $Po(\theta \log n)$, respectively, we have

$$\| \mathcal{L}(P_n|\theta) - \mathcal{L}(P_n^*|\theta) \| \leq \frac{\sqrt{\theta}|H_{\theta n} - \log n|}{\sqrt{H_{\theta n}} + \sqrt{\log n}}.$$

(For the upper bound of the total variation distance between two Poisson distributions, see Yannaros (1991), Theorem 2.1.) We note that

$$H_n - H_{\theta n} = \frac{1}{n} - \frac{1}{\theta} + \sum_{j=1}^{n-1} \frac{\theta}{j(\theta + j)}, \quad \sum_{j=1}^{n-1} \frac{\theta}{j(\theta + j)} \leq \frac{\pi^2}{6}\theta$$

and

$$\lim_{n \to \infty} (H_n - \log n) = C,$$

where $C$ is Euler's constant. Therefore, for sufficiently large $n$, we have

$$|H_{\theta n} - \log n| \leq |H_{\theta n} - H_n| + |H_n - \log n| \leq \frac{1}{\theta} + \frac{\pi^2}{6}\theta^2 + c_0$$

where $c_0$ ia a positive constant such that $c_0 > C + 1$. Therefore we have

$$\| \mathcal{L}(P_n|\theta) - \mathcal{L}(P_n^*|\theta) \| \quad \leq \quad \frac{1}{\sqrt{\log n}} \left( \frac{1}{\sqrt{\theta}} + \frac{\pi^2}{6}\theta^{3/2} + c_0\, \theta^{1/2} \right).$$

Thus, if $E_\gamma(\theta^{-1})$ and $E_\gamma(\theta^2)$ exist, then $E_\gamma(\theta^{-1/2})$, $E_\gamma(\theta^{1/2})$, $E_\gamma(\theta^{3/2})$ exist and we have

$$\| \mathcal{L}(P_n) - \mathcal{L}(P_n^*) \| = O\Big( \frac{1}{\sqrt{\log n}} \Big). \tag{6}$$

Therefore by (5) and (6) we have

$$\begin{aligned} \sup_{B \subset \mathbb{Z}_+} [P(K_n \in B) - P(P_n^* \in B)] &= \| \mathcal{L}(K_n) - \mathcal{L}(P_n^*) \| \\ &\leq \| \mathcal{L}(K_n) - \mathcal{L}(P_n) \| + \| \mathcal{L}(P_n) - \mathcal{L}(P_n^*) \| = O\Big( \frac{1}{\sqrt{\log n}} \Big). \end{aligned}$$

Thus, we have the following.

LEMMA 2.2. *We suppose that $E_\gamma(\theta^{-1})$ and $E_\gamma(\theta^2)$ exist. For $K_n$ and $P_n^*$, we have*

$$\sup_{-\infty < x < \infty} \left| P\Big( \frac{K_n}{\log n} \leq x \Big) - P\Big( \frac{P_n^*}{\log n} \leq x \Big) \right| = O\Big( \frac{1}{\sqrt{\log n}} \Big). \tag{7}$$

### 2.2. Poisson distribution, Normal approximation and mixture

Given $\theta$, since $P_n^*$ has the Poisson distribution $Po(\theta \log n)$, $(P_n^* - \theta \log n)/\sqrt{\theta \log n}$ converges to the standard normal distribution $N(0,1)$. Its order is given by

$$\sup_{-\infty<x<\infty} \left| P\left( \frac{P_n^* - \theta \log n}{\sqrt{\theta \log n}} \le x \;\middle|\; \theta \right) - \Phi(x) \right| \le \frac{0.8}{\sqrt{\theta \log n}}.$$

(For the upper bound of Normal approximation to Poisson distribution, see Michel (1993), Theorem 1.) Thus, given $\theta$, we have

$$\sup_{-\infty<x<\infty} \left| P\left( \frac{P_n^*}{\log n} \le x \middle| \theta \right) - \Phi\left( \frac{x-\theta}{\sqrt{\theta/\log n}} \right) \right| \le \frac{0.8}{\sqrt{\theta \log n}}.$$

Therefore, if $E_\gamma \theta^{-1/2} < \infty$, then we have

$$\begin{aligned} &\sup_{-\infty<x<\infty} \left| P\left( \frac{P_n^*}{\log n} \le x \right) - E_\gamma \Phi\left( \frac{x-\theta}{\sqrt{\theta/\log n}} \right) \right| \\ \le\;& E_\gamma \left\{ \sup_{-\infty<x<\infty} \left| P\left( \frac{P_n^*}{\log n} \le x \middle| \theta \right) - \Phi\left( \frac{x-\theta}{\sqrt{\theta/\log n}} \right) \right| \right\} = O\left( \frac{1}{\sqrt{\log n}} \right). \end{aligned} \tag{8}$$

We note that

$$E_\gamma \Phi\left( \frac{x-\theta}{\sqrt{\theta/\log n}} \right) \tag{9}$$

is the mixture distribution of the normal distribution $N(\theta, \theta/\log n)$ by $\theta$ having the distribution $\gamma$. Let $\varphi_0$ and $\varphi_\gamma$ be the characteristic functions of the distribution (9) and the distribution $\gamma$, respectively. Then we have

$$\begin{aligned} |\varphi_0(t) - \varphi_\gamma(t)| &= \left| E_\gamma e^{i\theta t - \frac{\theta}{2\log n} t^2} - E_\gamma e^{i\theta t} \right| = |E_\gamma e^{i\theta t}(e^{-\frac{\theta}{2\log n} t^2} - 1)| \\ &\le E_\gamma(1 - e^{-\frac{\theta}{2\log n} t^2}) \le E_\gamma(\theta) \frac{t^2}{2\log n} \end{aligned}$$

We see the difference between the distribution (9) and the distribution $\gamma$, by the smoothing lemma. (For the smoothing lemma, for example, see Feller (1966), p.538.) If the p.d.f. of the distribution $\gamma$ is bounded by $L(>0)$, then for any $\varepsilon > 0$ we have

$$\begin{aligned} \sup_x \left| E_\gamma \Phi\left( \frac{x-\theta}{\sqrt{\theta/\log n}} \right) - \gamma(x) \right| &\le \frac{1}{\pi} \int_{-\log^\varepsilon n}^{\log^\varepsilon n} \left| \frac{\varphi_0(t) - \varphi_\gamma(t)}{t} \right| dt + \frac{24L}{\pi \log^\varepsilon n} \\ &\le \frac{E_\gamma(\theta)}{2} \cdot \frac{1}{\log^{1-2\varepsilon} n} + \frac{24L}{\pi \log^\varepsilon n}. \end{aligned}$$

For the two terms on the right-hand side, the orders of $\log n$ coincide if and only if $\varepsilon = 1/3$, in which case the order of $\log n$ is $-1/3$. Therefore, we have

$$\sup_x \left| E_\gamma \Phi\left( \frac{x-\theta}{\sqrt{\theta/\log n}} \right) - \gamma(x) \right| = O\left( \frac{1}{\sqrt[3]{\log n}} \right). \tag{10}$$

Thus by (8) and (10) we have the following.

Lemma 2.3. *We suppose that p.d.f. of $\gamma$ is bounded, and that $E_\gamma(\theta^{-1/2})$ and $E_\gamma(\theta)$ exist. Then, for $P_n^*$ having the Poisson distribution $Po(\theta \log n)$ given $\theta$, we have*

$$\sup_{-\infty<x<\infty}\left|P\left(\frac{P_n^*}{\log n}\le x\right)-\gamma(x)\right| = O\left(\frac{1}{\sqrt[3]{\log n}}\right). \tag{11}$$

Since

$$\sup_{-\infty<x<\infty}\left|P\left(\frac{K_n}{\log n}\le x\right)-\gamma(x)\right|$$

$$\le \sup_{-\infty<x<\infty}\left|P\left(\frac{K_n}{\log n}\le x\right)-P\left(\frac{P_n^*}{\log n}\le x\right)\right| + \sup_{-\infty<x<\infty}\left|P\left(\frac{P_n^*}{\log n}\le x\right)-\gamma(x)\right|$$

by (7) and (11), we get Proposition 2.1.

At last, we note about the assumption of Propsition 2.1 that p.d.f. of $\gamma$ is bounded, and that $E_\gamma(\theta^{-1})$ and $E_\gamma(\theta^2)$ exist. (I) For the Rayleigh distribution whose p.d.f. is given by $g(x) = (x/b^2)\exp(-x^2/2b^2)$ $(x > 0;\ b > 0)$ , the assumption is satisfied. (II) For the gamma distribution whose p.d.f. is given by $g(x) = (x/b)^{c-1}e^{-x/b}/b\Gamma(c)$ $(x > 0;\ b, c > 0)$ , the assumption is satisfied in case of $c > 1$. (III) For the triangular distribution whose p.d.f. is given by $g(x) = 2x/bc$ $(0 < x \le c)$ and $2(b-x)/[b(b-c)]$ $(c < x < b)$ for $0 < c < b$ , the assumption is satisfied.

The rate of convergence given by (3) depends on (10), which is derived by using the smoothing lemma. For the better rate, the evaluation of the left-hand side of (10) must be improved. Further work is necessary on the evaluation of the left-hand side of (10) and the convergence of $K_n/\log n$ $(n \to \infty)$.

**Acknowledgement**

The author is grateful to the referee for his careful reading and useful comments. This work was supported by Grant-in-Aid for Scientific Research (B) (No. 22300097), Japan Society for the Promotion of Science.

## References

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

Barbour, A.D. and Hall, P. (1984). On the rate of Poisson convergence. *Math. Proc. Camb. Phil. Soc.*, **95**, 473–480.

Feller, W. (1970). *An introduction to probability theory and its applications Vol. II.* New York : John Wiley & Sons.

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

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

Michel, R (1993). On Berry-Esseen bound for the compound Poisson distribution. *Insurance: Mathematics and Economics*, **13**, 35–37.

Sethuraman, J (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, **4**, 639–650.

Yannaros, N. (1991). Poisson approximation for random sums of Bernoulli random variables. *Statistics and Probability Letters*, **11**, 161–165

*Received April 2, 2012*
*Revised September 21, 2012*