Limit Theorems for the Frequency Counts of the Ewens-Pitman Model

Shui Feng
McMaster University

10th Workshop on Markov Processes and Related Topics
Xian University and Beijing Normal University, Xian, China
August 14-20 , 2014

- Random Partitions

- Construction Through Random Sampling

- Unconditional Results
  - Total Frequency Counts
  - Frequency Counts

- Conditional Results
  - Total Frequency Counts
  - Frequency Counts

Random Partitions

For any integer $n \geq 1$ and $1 \leq k \leq n$, a partition π of the set $[n] = \{1, \ldots, n\}$ into k blocks is a collection of k nonempty unordered disjoint subsets $A_1, \ldots, A_k$ of $\{1, \ldots, n\}$ such that

$$\{1, \ldots, n\} = \cup_{i=1}^{k} A_i.$$

Let $\mathcal{P}_n$ be the set of all finite partitions of $\{1, \ldots, n\}$.

Definition: A *random partition* is a random variable Π_n taking values in $\mathcal{P}_n$. The partition Π_n is *exchangeable* if its law is invariant under permutations.

Definition: A consistent (in terms of restriction) family of $\{\Pi_n : n \geq 1\}$ denoted by Π is called a random partition of $\mathbb{N} = \{1, 2, \ldots\}$.

Definition: A random partition Π is *exchangeable* if for each $n \geq 1$ Π_n is exchangeable.

Let $|A|$ denote the number of elements in A and set

$$n_i = |A_i|,\ i = 1, \dots, k.$$

Then $n_1, \dots, n_k$ is clearly a partition of integer n. For any $1 \le j \le n$, set

$$m_j = \#\{1 \le i \le k : n_i = j\}.$$

Clearly

$$\sum_{j=1}^{n} j m_j = n,\ \sum_{j=1}^{n} m_j = k.$$

A simple example: $n = 12$ and

$$\{1,2,3,4,5,6,7,8,9,10,11,12\}$$
$$= \{1,7,9\} \cup \{2,6\} \cup \{3,4,5\} \cup \{8,10,12\} \cup \{11\}.$$

For this particular partition, we have

$$k = 5$$
$$m_1 = 1, m_2 = 1, m_3 = 3, m_4 = \cdots = m_{12} = 0.$$

The total number of set partitions of $\{1, \ldots, n\}$ corresponding to $m_1, \ldots, m_n$ is

$$D(m_1, \ldots, m_n) = \frac{n!}{\prod_{j=1}^{n} (j!)^{m_j} m_j!}.$$

Given a random partition Π_n, let K_n be the total number of sets in the partition, $N_1^n, \ldots, N_{K_n}^n$ the corresponding set sizes, and

$$M_j^n = \#\{i : N_i^n = j\}.$$

Definition: $M_1^n, \ldots, M_n^n$ are called the *frequency counts* of the random partition Π_n.

$K_n = \sum_{j=1}^{n} M_j^n$ is the *total frequency counts*.

Construction of Random Partitions

– Poisson point process
– Subordinator and excursion
– Gnedin's random open sets
– Kingman's paintbox
– Random distributions
– Species sampling
– Urn models (e.g. Chinese restaurant process)

Construction Through Random Sampling

Let $(X_1, \ldots, X_n)$ be a random sample of size n from a population following certain "nice "distribution.

A random partition is constructed so that the i, j belongs to the same family iff $X_i = X_j$.

Total frequency count $K_n =$ the number of distinct families in the sample.

Frequency count $M_j^n =$ number of families of size j.

Given $X_1, \ldots, X_n$, additional samples of size m are selected resulting in a sample of total size $n+m$: $X_1, \ldots, X_n, X_{n+1}, \ldots, X_{n+m}$.

Let $K_m^{(n)} = K_{m+n} - K_n$ denote the total frequency counts of new blocks introduced by the additional sample of size m.

Questions

1 What happens to K_n and M_j^n for large n? (unconditional setting)

2 Given $X_1, \ldots, X_n$, what happens to $K_m^{(n)}$ and $M_j^{(n+m)}$ for large m? (conditional setting)

Example Consider a population of r types of individuals with corresponding proportions $p_1, \ldots, p_r$. Taking a random sample $X_1, \ldots, X_n$ from the population and introducing the equivalent relation $i \sim j$ iff $X_i = X_j$. This leads to a random partition of $\{1, 2, \ldots, n\}$.

Possible generalizations:

1 The number of types r becomes infinity.

2 $p_1, \ldots, p_r$ becomes random.

The random sample $X_1, \ldots, X_n$ will be exchangeable instead of iid.

3 Both 1 and 2.

Finite r

A nice randomization of $p_1, \ldots, p_r$ is the Dirichlet distribution.

$r = \infty$

A nice choice would be the so-called two-parameter Poisson-Dirichlet distribution or equivalently

$$p_1 = U_1,\ p_n = (1 - U_1) \cdots (1 - U_{n-1}) U_n, n \geq 2$$

where $U_1, U_2, \ldots$ are independent Beta random variables with U_i following the $beta(1 - \alpha, \theta + i\alpha)$ distribution for some $0 < \alpha < 1, \theta + \alpha > 0$.

The latter is the focus of this talk.

Pitman Sampling Formula

For each $m_1, \ldots, m_n$, and $0 < \alpha < 1, \theta > -\alpha$

$$\mathbb{P}\{M_j^n = m_j, j = 1, \ldots, n\} = D(m_1, \ldots, m_n) \frac{(\theta)_{k \uparrow \alpha}}{(\theta)_{n \uparrow 1}} \prod_{i=1}^{n} [(1-\alpha)_{i \uparrow 1}]^{m_i},$$

where the notation

$$(a)_{n \uparrow b} = a(a+b) \cdots (a+(n-1)b).$$

This is the two-parameter Pitman model or the Ewens-Pitman model.

Unconditional Results

Total Frequency Counts K_n

The total frequency counts $\{K_n\}_{n \geq 1}$ is a nondecreasing Markov chain with $K_1 = 1$ and for any $k \geq 1$

$$\begin{aligned}
\mathbb{P}\{K_{n+1} = k+1 | K_1, \ldots, K_n = k\} &= \frac{k\alpha + \theta}{n + \theta} \\
\mathbb{P}\{K_{n+1} = k | K_1, \ldots, K_n = k\} &= \frac{n - k\alpha}{n + \theta}.
\end{aligned}$$

This describes a natural urn structure as follows.

- Consider an urn that initially contains a black ball of mass θ.
- Balls are drawn from the urn successively with probabilities proportional to their masses.
- When a black ball is drawn, it is returned to the urn together with a black ball of mass α and a ball of new colour with mass $1-\alpha$.
- If a non-black ball is drawn, it is returned to the urn with one additional ball of mass one with the same colour.
- Colors are labelled $1, 2, 3, ...$ in the order of appearance.

The total frequency counts represent the total number of different new colours after the nth draw.

Given $0 < \alpha < 1, \theta > -\alpha$, the distribution of K_n is given by

$$\mathbb{P}\{K_n = k\} = \frac{(\theta+\alpha)_{k-1\uparrow\alpha}}{(\theta+1)_{n-1\uparrow 1}} S_\alpha(n,k)$$

where $S_\alpha(n,k)$ is a generalized Stirling number of the first kind satisfying

$$x(x+1)\cdots(x+n-1) = \sum_{i=0}^{n} S_\alpha(n,i)x(x+\alpha)\cdots(x+(i-1)\alpha)$$

or equivalently

$$(x)_{n\uparrow 1} = \sum_{i=0}^{n} S_\alpha(n,i)(x)_{i\uparrow\alpha}.$$

Fluctuation

Let $S_{\alpha,\theta}$ be a positive continuous random variable with density function

$$g_{\alpha,\theta}(x) = \frac{\Gamma(\theta+1)}{\Gamma(\frac{\theta}{\alpha}+1)} x^{\frac{\theta}{\alpha}} g_\alpha(x),$$

where

$$g_\alpha(x) = \frac{1}{\pi\alpha} \sum_{i=0}^{\infty} \frac{(-1)^{i+1}}{i!} \Gamma(i\alpha+1) x^{i-1} \sin(\pi\alpha i)$$

is the density function of the Mittag-Leffler distribution.

Theorem 1. (Pitman (97))

$$\lim_{n\to\infty} \frac{K_n}{n^\alpha} = S_{\alpha,\theta} \ a.s.$$

Large Deviations

Define

$$\Lambda_\alpha(\lambda) = \begin{cases} -\log[1-(1-e^{-\lambda})^{\frac{1}{\alpha}}] & \text{if } \lambda > 0, \\ 0, & \text{else} \end{cases}$$

and

$$I^\alpha(x) = \sup_\lambda \{\lambda x - \Lambda_\alpha(\lambda)\},$$

Theorem 2. (F and Hoppe(98)) *For appropriate subset A of $[0,\infty)$,*

$$\mathbb{P}\{K_n/n \in A\} \asymp \exp\{-n \inf_{x\in A} I^\alpha(x)\}.$$

Key Calculations in the Proof

For $\theta = 0$,

$$\mathbb{E}[(K_n)^i] = \frac{\Gamma(i)(\alpha i)_{n\uparrow 1}}{\alpha\Gamma(n)}.$$

For $\lambda >$ and $x = 1 - e^{-\lambda}$,

$$\mathbb{E}[(\frac{1}{1-x})^{K_n}] = \sum_{i=0}^{\infty} x^i \binom{\imath + n - 1}{n-1}.$$

This leads to

$$\frac{1}{n}\log\mathbb{E}[e^{\lambda K_n}] \to \Lambda_\alpha(\lambda).$$

Frequency Counts M_j^n

Lemma 3. (Favaro and F (2014a)) *For any integers* $j, r \geq 1$*, we have for* $\theta \neq 0$

$$\mathbb{E}[(M_j^n)_{r\uparrow 1}]$$
$$= \frac{1}{(\theta)_{n\uparrow 1}} \sum_{i=0}^{r} \binom{r-1}{r-i} \frac{r!}{i!} \left(\alpha \frac{(1-\alpha)_{(l-1)\uparrow 1}}{l!} \right)^i \left(\frac{\theta}{\alpha} \right)_{i\uparrow 1} (n)_{il\downarrow 1} (\theta + i\alpha)_{(n-il)\uparrow 1}$$

and for $\theta = 0$*,*

$$\mathbb{E}[(M_j^n)_{r\uparrow 1}]$$
$$= \frac{1}{\alpha \Gamma(n)} \sum_{i=0}^{r} \binom{r}{i} (r-1)! \left(\alpha \frac{(1-\alpha)_{(l-1)\uparrow 1}}{l!} \right)^i (n)_{il\downarrow 1} (i\alpha)_{(n-il)\uparrow 1}.$$

In particular, one has

$$\mathbb{E}[M_j^n] = \frac{\Gamma(\theta+1)}{\Gamma(\theta+\alpha)} \frac{(1-\alpha)_{(j-1)\uparrow 1}}{j!} \frac{\Gamma(\theta+\alpha+n-j)}{\Gamma(\theta+n)} \frac{n!}{(n-j)!}$$

$$\begin{aligned} &\mathbb{E}[M_j^n(M_j^n-1)] \\ &= \frac{(\theta+\alpha)\Gamma(\theta+1)}{\Gamma(\theta+2\alpha)} (\frac{(1-\alpha)_{(j-1)\uparrow 1}}{j!})^2 \frac{\Gamma(\theta+2\alpha+n-2j)}{\Gamma(\theta+n)} \frac{n!}{(n-2j)!}. \end{aligned}$$

Let n tends to infinity, we obtain

$$\mathbb{E}[\frac{M_j^n}{n^\alpha}] \to \frac{\Gamma(\theta+1)}{\Gamma(\theta+\alpha)}\frac{(1-\alpha)_{(j-1)\uparrow 1}}{j!},$$

$$\text{Var}[\frac{M_j^n}{n^\alpha}] \to \Gamma(\theta+1)\{\frac{(\theta+\alpha)}{\Gamma(\theta+\alpha)} - \frac{\Gamma(\theta+1)}{[\Gamma(\theta+\alpha)]^2}\}(\frac{(1-\alpha)_{j-1\uparrow 1}}{j!})^2.$$

Define

$$S_{\alpha,\theta,j} = \frac{\alpha\Gamma(j-\alpha)}{\Gamma(1-\alpha)\Gamma(j+1)}S_{\alpha,\theta}, j = 1, 2, \ldots.$$

LLN

For any $j \geq 1$,

$$\frac{M_j^n}{n} \to 0, n \to \infty, \quad a.s$$

Fluctuation(Pitman(97)):

$$\frac{M_j^n}{n^\alpha} \Rightarrow S_{\alpha,\theta,j}, \quad n \to \infty.$$

Large Deviations

For $\lambda > 0$, let $x = 1 - e^{-\lambda}$ and

$$F_n(x;\theta,\alpha) = \mathbb{E}[e^{\lambda M_j^n}] = \mathbb{E}\left[\left(\frac{1}{1-x}\right)^{M_j^n}\right].$$

Theorem 4.

$$F_n(x;\theta,\alpha)$$

$$= \frac{1}{(\theta)_{n\uparrow 1}} \sum_{i=0}^{\lfloor \frac{n}{j} \rfloor} \left(\frac{x}{1-x}\right)^i \left(\alpha \frac{(1-\alpha)_{(j-1)\uparrow 1}}{j!}\right)^i \frac{1}{i!} \left(\frac{\theta}{\alpha}\right)_{i\uparrow 1} (n)_{ij\downarrow 1} (\theta + i\alpha)_{(n-ij)\uparrow 1}.$$

In particular for $\theta = 0$, we have

$$F_n(x;0,\alpha) = \sum_{i=0}^{\lfloor \frac{n}{j} \rfloor} \left(\frac{x}{1-x}\right)^i \left(\alpha \frac{(1-\alpha)_{(j-1)\uparrow 1}}{j!}\right)^i \frac{n}{n-ij} \binom{n-ij+i\alpha-1}{n-ij-1}.$$

For $\lambda \le 0$, set $\Lambda_{\alpha,j} = 0$. For $\lambda > 0$, let

$$\tilde{x} = \frac{\alpha x (1-\alpha)_{(j-1)\uparrow 1}}{(1-x) j!}$$

and $\varepsilon_0(\lambda)$ be the unique solution of the equation

$$(j-\alpha)\log(1-(j-\alpha)\varepsilon) - j\log(1-j\varepsilon) - \alpha\log\alpha\varepsilon - \log\tilde{x} = 0.$$

For $\lambda > 0$, define

$$\Lambda_{\alpha,j}(\lambda) = \log[1 + \frac{\alpha\varepsilon_0}{1 - j\varepsilon_0}]$$

and

$$I_j(y) = \sup\{\lambda y - \Lambda_{\alpha,j}(\lambda) : \lambda \in \mathbb{R}\}.$$

Theorem 5. (Favaro and F(2014b)) *For any measurable set* $A \subset \mathbb{R}$, *set* $I_j(A) = \inf\{I_j(y) : y \in A\}$. *Then*

$$-I_j(A^\circ) \leq \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\{\frac{M_j^n}{n} \in A\} \leq \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\{\frac{M_j^n}{n} \in A\} \leq -I_j(\bar{A})$$

where A° *and* $\bar{A}$ *are the interior and closure of* A *respectively. In other words the family* $\{M_j^n/n : n \geq 1\}$ *satisfies a LDP under the two-parameter Dirichlet process with good rate function* $I_j(\cdot)$ *as* n *tends to infinity.*

Conditional Results

Total Frequency Counts

Let $\mathbb{P}_l$ and $\mathbb{E}_l$ denote the respective conditional law and conditional expectation given $K_n = l$.

Theorem 6. (Favaro and F (2014a)) *For* $\lambda > 0$*, let* $x = 1 - e^{-\lambda}$*. Then*

$$\mathbb{E}_l[e^{\lambda K_m^{(n)}}]$$

$$= (1-x)^{l+\frac{\theta}{\alpha}} \sum_{k \geq 0} \frac{x^k}{k!} (l + \frac{\theta}{\alpha})_{k \uparrow 1} \frac{\binom{n+\theta+k\alpha+m-1}{n+\theta+m-1}}{\binom{n+\theta+k\alpha-1}{n+\theta-1}}.$$

Fluctuation

For any $c, d > 0$, let $B_{c,d}$ denote the beta random variable with parameters c and d. Let $S^{l,n}_{\alpha,\theta}$ be the independent product of a beta random variable $B_{l+\theta/\alpha,n/\alpha-l}$ and $S_{\alpha,\theta}$.

Theorem 7. (Favaro et al (2009)) *Under $\mathbb{P}_l$, we have*

$$\frac{K^{(n)}_m}{m^\alpha} \to S^{l,n}_{\alpha,\theta} \;\; a.s. \text{ as } m \to \infty.$$

In other words, the condition on the first n samples has a long lasting impact about the fluctuation of future samples.

Large Deviations

Theorem 8. (Favaro and F(2014a)) *For any measurable set $A \subset \mathbb{R}$,*

$$-I(A^\circ) \leq \liminf_{m\to\infty} \frac{1}{m} \log \mathbb{P}_l\{\frac{K_m^{(n)}}{m} \in A\} \leq \limsup_{m\to\infty} \frac{1}{m} \log \mathbb{P}_l\{\frac{K_m^{(n)}}{m} \in A\} \leq -I(\bar{A})$$

where A° and $\bar{A}$ are the interior and closure of A respectively, and the rate function is the same as the unconditional case.

Frequency Counts

Fluctuation

Theorem 9. (Favaro et al (2009)) *Under $\mathbb{P}_l$, we have*

$$\frac{M_j^{(n+m)}}{m^\alpha} \to \frac{\alpha\Gamma(j-\alpha)}{\Gamma(1-\alpha)\Gamma(j+1)} S_{\alpha,\theta,}^{l,n} \quad a.s. \text{ as } m \to \infty.$$

In other words, the condition on the first n samples has a long lasting impact about the fluctuation of future samples.

Theorem 10. (Favaro and F(2014b)) *For any measurable set* $A \subset \mathbb{R}$*,*

$$\begin{aligned} -I_j(A^\circ) &\le \liminf_{m\to\infty} \frac{1}{m} \log \mathbb{P}_l\{\frac{M_j^{(n+m)}}{m} \in A\} \\ &\le \limsup_{m\to\infty} \frac{1}{m} \log \mathbb{P}_l\{\frac{M_j^{(n+m)}}{m} \in A\} \\ &\le -I_j(\bar{A}) \end{aligned}$$

where the rate function turns out to be the same as the unconditional case.

References

S. Favaro and S. Feng (2014a). Asymptotics for the number of boocks in a conditional Ewens-Pitman sampling models. *Electron. J. Probab.*, Vol 19, 1–15.

S. Favaro and S. Feng (2014b). Large sample asymptotics of the frequency counts associated with the two-parameter Dirichlet prior and posterior. *preprint*.

S. Favaro, A. Lijoi and I. Prünster (2009). Bayesian nonparametric inference for species variety with a two parameter Poisson-Dirichlet process prior. *J. Roy. Statist. Soc. Ser. B.*, **71**, 993–1008.

S. Feng and F.M. Hoppe (1998). Large deviation principles for some random combinatorial structures in populartion genetics and Brownian motion. *Ann. Appl. Probab.*, **8**, 975–994.

J. Pitman (1997). Partition structures derived from Brownian motion and stable subordinators. *Bernoulli*, **3**, 79–96.