Asymptotic Moments Matching to Uniformly Minimum Variance Unbiased Estimation under Ewens Sampling Formula

Masayo Y. Hirose*  Shuhei Mano†

Abstract

The Ewens sampling formula is a distribution related to the random partition of a positive integer. In this study, we investigate the issue of non-existence solutions in parameter estimation under the distribution. As a result, the first and second moments matching estimators to the uniformly minimum variance unbiased estimator are derived using the Ewens sampling formula in asymptotic sense. A Monte Carlo simulation study is performed to evaluate the efficiency of the resulting estimators.

Keywords: Ewens Distribution, Higher-Order Efficiency, Population Unique.

1 Introduction

Ewens (1972) provided the Ewens sampling formula, which is a law of the partition of positive integers into components comprising non-negative integers in the context of genetics. Antoniak (1974) derived the Ewens sampling formula in the context of Bayesian statistics as a partition induced by a sample from the Dirichlet process. It has been applied to several research fields in ecology, disclosure control and so on. Let $S \in S_N$ satisfy

$$S_N = \{S \equiv (S_1, \ldots, S_N) : S_i \geq 0, \sum_{i=1}^{N} iS_i = N, i \in \{1, \ldots, N\}\}.$$ 

*Institute of Mathematics for Industry, Kyushu University, masayo@imi.kyushu-u.ac.jp
†Department of Statistical Inference and Mathematics, The Institute of Statistical Mathematics, smano@ism.ac.jp
Elements $S_i$ of the $N$-dimensional vector $S$ are random variables and denote the number of types of $i$ times appear. This is known as the frequency of the frequencies (Good, 1953). Next, we define $K_N = \sum_{i=1}^{N} S_i$. The number of types to appear is denoted, i.e., the length of the random partition. For instance, in ecology, $S_i$ and $K_N$ denote the number of species that appear $i$ times and the number of different species to occur, respectively. For a positive integer $N$, it is expressed using the parameter $\theta > 0$ as follows:

$$P(S = (s_1, \ldots, s_N) \in \mathbb{S}_N) = \frac{N!\theta^k}{\theta^{[N]} \prod_{j=1}^{N} j^{s_j} s_j!},$$  \hspace{1cm} (1)$$

where $k = \sum_{i=1}^{N} s_i$ and $\theta^{[N]} = \prod_{j=1}^{N} (\theta + j - 1)$.

The parameter $\theta$ controls the "diversity": $k \to N$ as $\theta \to \infty$, and $k \to 1$ as $\theta \to 0$. It is of particular interest to assess $S_1$. For instance, $S_1$ denotes the number of singleton species in the population and a population unique. The latter is explained in Section 4 as an example. For more details regarding this formula, readers may refer to Tavare and Ewens (1997), Crane (2016), and Mano (2018).

The distribution of $K_N = \sum_{i=1}^{N} S_i$, given $N$, can be obtained as follows:

$$P(K_N = k) = \frac{\theta^k}{\theta^{[N]} s(N, k)},$$  \hspace{1cm} (2)$$

where $s(N, k)$ is the unsigned Stirling numbers of the first kind satisfying $\theta^{[N]} = \sum_{k=1}^{N} s(N, k)\theta^k$ for non-negative integers $k$ and $N$ such that $1 \leq k \leq N < \infty$.

We let $R_i = E[S_i]$ be the expectation of $S_i$ with respect to this model. Its explicit formula is obtained as a function of $\theta$ in Watterson (1974) as follows:

$$R_i = R_i(\theta) = \frac{\theta}{i} \prod_{j=1}^{N} \frac{N-j+1}{\theta + N - j}.$$
any subsampling partitions with sample size $n$ from population size $N$ for all $n < N$. Therefore, in practice, for sample size $n$, an estimator of $R_i$ can be obtained by replacing the unknown parameter $\theta$ in the formula of $R_i$ with its consistent estimator.

In particular, the maximum likelihood estimator $\hat{\theta}_{ML}$ is widely used and is obtained as follows:

$$\hat{\theta}_{ML} = \arg \max_{\theta \in \mathbb{R}_{>0}} l(\theta),$$

where $\mathbb{R}_{>0}$ is the parameter space of $\theta$, and the logarithm of the likelihood $l(\theta)$ is such that

$$l(\theta) = K_n \log(\theta) - \sum_{j=1}^{n} \log(\theta + j - 1) + \text{const.}$$

Tavare and Ewens (1997) reported, “This estimator is biased, but the bias decreases as $n$ increases.” In addition, if $R_i$ is U-estimable, it may be difficult to obtain a uniformly minimum variance unbiased estimator (UMVUE). To the best of our knowledge, asymptotic moments matching estimators to a uniformly minimum variance unbiased estimator of $R_i$ have not been investigated, whereas $E[K_n]$ is an exact UMVUE. We denote this estimator the asymptotic UMVUE hereinafter. Moreover, even if we can identify such an estimator, it may yield unrealistic negative estimates of non-negative $R_i$.

First, we address the construction of the asymptotic UMVUE of $R_i$ with a small positive integer $i$, up to the second moment matching in second-order asymptotic sense. In this study, the second-order indicates the order $O((\log n)^{-1})$ for large $n$. We also avoid severe issues such that the maximum likelihood solution could not exist (see Section 2). In addition, the precision of the estimation can be improved further. Second, we construct a higher-order asymptotic UMVUE by matching the first and second moments.

For the second purpose, we use two types of bias correction methods: additive bias correction and the adjusted maximum likelihood method. The adjusted maximum likelihood is multiplied a nonrandom adjustment factor by the likelihood. It was developed by Lahiri and Li (2010), Li and Lahiri (2010), and Yoshimori and Lahiri (2014) to avoid zero estimates of dispersion parameters in a linear mixed model, particularly in the research fields of small-area estimation. Hirose and Lahiri (2018) achieved a second-order asymptotic unbiasedness of several important re-parameterized estimators by suggesting a new adjusted maximum likelihood method. Hirose and
Mano (2021) constructed a general framework using differential geometry to achieve second-order unbiasedness and applied the methodology to a general model with multi-dimensional parameters.

One may consider these previous results to be applicable because the Ewens distribution is an exponential distribution family. Nevertheless, these results are insufficient to obtain the asymptotic UMVUE up to the fourth-order asymptotic sense, unlike the studies of Hirose and Lahiri (2018) and Hirose and Mano (2021). It is note that, in this study, the second (or fourth) order denotes the order of $O((\log n)^{-1})$ (or $O((\log n)^{-2})$) for large $n$, whereas the second-order denotes the $O(n^{-1})$ in the study of Hirose and Mano (2021), considering the differences in the Fisher information order.

As a result, it is sufficient to obtain one common estimate of the parameter for achieving the second purpose even when $R_i$ and $R_j$ are to be estimated simultaneously for $i \neq j$. Furthermore, we demonstrate the higher-order asymptotic results based on easier proofs, owing to the functional form $R_i$ of $\theta$, a property of the exponential distribution family and a relationship between the parameter $\theta$ and natural parameter $\xi$. For more details, see Section 3.2 and Appendix C.3.

The remainder of this paper is organized as follows: In Section 2, we introduce the existing estimator of $R_i$ and modify it to be the asymptotic UMVUE, up to the second-order, to achieve the first purpose. The problem of non-existence of estimates is avoided in this section. To address the second problem, higher-order asymptotic unbiasedness is discussed in Section 3. In this section, we suggest two estimators using two bias-correction methods. This methodology can be applied to practical issues. Subsequently, in Section 4, we present an example where our methodology is applied to estimate the number of population uniques to disclosure control. The simulation study is described in Section 5. Herein, it is assumed that $\theta$ and $i$ are bounded for large $n$. In addition, we assume that the sampling design is simple random sampling without replacement. All technical proofs are provided in the appendix.

2 Maximum Likelihood Estimation of Parameter

As mentioned earlier, a typical method to estimate $R_i$ is to replace the $\theta$ of $R_i$ with its maximum likelihood estimator. Its first and second derivatives

\[ \frac{\partial^2 R_i}{\partial \theta^2} = \frac{1}{n} \left( \frac{1}{\theta^2} + \frac{1}{(\log n)^2} \right) \]

\[ \frac{\partial^3 R_i}{\partial \theta^3} = \frac{1}{n} \left( \frac{2}{\theta^3} + \frac{6}{\theta \log n} + \frac{2}{(\log n)^3} \right) \]
are expressed as follows:

\[
\frac{\partial l(\theta)}{\partial \theta} \equiv \partial_{\theta} l(\theta) = \frac{K_n}{\theta} - \sum_{j=1}^{n} \frac{1}{\theta + j - 1}, \tag{3}
\]

\[
\frac{\partial^2 l(\theta)}{\partial \theta^2} \equiv \partial_{\theta}^2 l(\theta) = - \frac{K_n}{\theta^2} + \sum_{j=1}^{n} \frac{1}{(\theta + j - 1)^2}. \tag{4}
\]

From (3), the maximum likelihood estimator \( \hat{\theta}_{ML} \) can be expressed as the root of

\[
\frac{K_n}{\theta} - \sum_{j=1}^{n} \frac{1}{\theta + j - 1} = 0.
\]

We let \( \hat{R}^{(N)}_i = R_i(\hat{\theta}_{ML}) \) and refer to it as a naive estimator.

The continuous mapping theorem may provide the consistency of the naive estimator; to the best of our knowledge, the properties of asymptotic moments have not been investigated hitherto.

The following practical issue occurs when attempting to obtain an asymptotic UMVUE: cases \( K_n \in \{1, n\} \) provide each likelihood as a strict monotone function of \( \theta \). Additionally, it is shown from the first-order derivative of the likelihood function of the natural parameter \( \xi = \log \theta \in \Xi \subset \mathbb{R} \) on the Ewens distribution.

\[
\frac{\partial \xi l(\xi)}{\partial \xi} = K_n - n + \sum_{j=2}^{n} \frac{j - 1}{e^\xi + j - 1}.
\]

It is note that the equation above is rewritten from \( \text{[3]} \) with the natural parameter \( \xi \).

The following lemma is established to assess such probability.

**Lemma 1.** Under the regularity condition \( R1 \), we have the following for large \( n \):

\[
P(\mathcal{K}) = o((\log n)^{-2}),
\]

where the set \( \mathcal{K} = \{ K_n \in \{1, n\} \} \).

The regularity condition and proof are provided in Appendices A and C.1, respectively.
In set $\mathcal{K}$, the lemma implies the problem of non-existing maximum likelihood solutions in $\Theta \subset \mathbb{R}_{>0}$ with extremely low but non-zero probability. In addition, $\hat{R}_i^{(N)}$ may be absent.

Therefore, the estimator must be modified to obtain an asymptotic UMVUE by considering such cases. Hence, we define set $\mathcal{S} = \{\hat{\theta} : \hat{\theta} \in (0, C_+)]$ with a large positive finite value $C_+$, which does not depend on $n$. Subsequently, we let

$$
\hat{R}_i^{(NM)} = \begin{cases} 
0 & (K_n = 1) \\
R_i(\hat{\theta}_{ML}) & (\hat{\theta}_{ML} \in \mathcal{S}) \\
R_i(C_+) & (\hat{\theta}_{ML} \notin \mathcal{S} \cup \{0\}).
\end{cases}
$$

It is note that $R_i(C_+)$ can be adopted in case $\{K_n = n\}$ because $\{K_n = n\} \subset \mathcal{S}^c \cap \{0\}^c$. In addition, $\hat{\theta}_{ML} > 0$ when $1 < K_n < n$ because $\xi$ exists such that $\partial l(\xi) = 0$.

Next, another lemma is established, the proof of which is provided in Appendix C.2.

**Lemma 2.** Under regularity conditions R1 and R2, we have the following for large $n$:

$$P(\mathcal{S}^c) = o((\log n)^{-2}).$$

Therefore, the estimator $\hat{R}_i^{(NM)}$ is the function of the complete sufficient statistic $K_n$ of $\theta$ in cases $\hat{\theta}_{ML} \in \mathcal{S}$. From Lemmas 1 and 2, Theorem 1 shows that estimator $\hat{R}_i^{(NM)}$ is the second-order asymptotic UMVUE for large $n$.

**Theorem 1.** Under regularity condition R1, the following holds:

$$(i) E[\hat{R}_i^{(NM)} - R_i(\theta)] = O((\log n)^{-2}),$$

$$(ii) E[(\hat{R}_i^{(NM)} - R_i(\theta))^2] = \frac{R_i^2}{\sum_{j=2}^{\theta} (\theta+1)^2} + o((\log n)^{-2}).$$

This is shown in Appendix B.1.

Next, we provide a remark.

**Remark 2.1.** The length of the random partition $K_N$ in the population is also of interest to infer. It holds that

$$\eta(\theta) = E[K_N] = \sum_{j=1}^{\theta} \theta \frac{1}{\theta + j - 1}.$$
We do not address improving this estimator because the estimator \( \eta(\hat{\theta}_{ML}) \) becomes the exact UMVUE of \( \eta(\theta) \). The result is a well-known result and can be an example of Corollary 3.13 in Hirose and Mano (2021).

3 Higher-Order Asymptotic UMVUE

3.1 General bias-corrected estimator

Theorem 1 shows that the bias is of the order \( O((\log n)^{-2}) \) for large \( n \). In this section, we address the construction of two types of asymptotic UMVUEs for matching the first and second moments in the fourth-order asymptotic sense for large \( n \). In Sections 3.2 and 3.3, we present the results using additive bias correction and the adjusted maximum likelihood method, respectively.

Let \( \hat{R}_i^{BC} \) denote the general bias-corrected estimator of \( R_i \) while considering set \( \mathcal{K} \).

\[
\hat{R}_i^{(BC)} = \begin{cases} 
0 & (K_n = 1) \\
\hat{R}_i(\hat{\theta}) & (\hat{\theta} \in \mathcal{S}) \\
R_i(C_+) & (\hat{\theta} \notin \mathcal{S} \cup \{0\}),
\end{cases}
\]  

provided \( \hat{\theta} > 0 \), where \( \hat{R}_i(\hat{\theta}) \) and \( \hat{\theta} \) denote consistent estimators of \( R_i \) and \( \theta \) for large \( n \), respectively. We set \( \hat{\theta} \) as a small positive value when the solution of \( \theta \) is negative, except for \( K_n = 1 \). However, it holds that \( \hat{\theta}_{ML} > 0 \) when \( 1 < K_n < n \).

3.2 Additive bias correction for the higher-order asymptotic inference

One may consider using the additive bias correction method to reduce bias. Let \( \hat{R}_i^{(BC1)} \) be the term \( \hat{R}_i(\hat{\theta}) \) in (5) replaced with \( \hat{R}_i^{(N)} - B_i(\hat{\theta}_{ML}) \), where

\[
B_i(\theta) = \frac{\sum_{j=2}^{n} \frac{(j-1)}{(\theta+j-1)^2} R_i}{\left\{ \sum_{j=2}^{n} \frac{(j-1)}{(\theta+j-1)^2} \right\}^2}
\]

Then, the theorem establishes that \( \hat{R}_i^{(BC1)} \) achieves the fourth-order unbiasedness for large \( n \), while maintaining the asymptotic efficiency.
Theorem 2. Under regularity condition R1 for large n, the following holds:

(i) \( E[\hat{R}_i^{(BC1)} - R_i(\theta)] = o((\log n)^{-2}) \),

(ii) \( E[(\hat{R}_i^{(BC1)} - R_i(\theta))^2] = R_i^2 \sum_{j=2}^{n} \frac{\theta(j-1)}{(\theta+j-1)^2} + o((\log n)^{-2}) \).

The proof is provided in Appendix B.1.

In addition, the estimator \( \hat{R}_i^{(BC1)} \) is a function of the complete sufficient statistic \( K_n \) of \( \theta \) in cases \( \hat{\theta}_{ML} \in S \). Therefore Lemmas [1, 2] and Theorem [2] prove that it is the asymptotic UMVUE, up to the fourth-order.

3.3 Another possible bias correction: adjusted maximum likelihood method

Alternatively, bias can also be reduced using the adjusted maximum likelihood method. This method has been developed by Hirose and Lahiri (2018) and Hirose and Mano (2021) for bias correction after re-parameterization. Firth (1993) suggested a similar bias reduction method for \( \hat{\theta}_{ML} \) via second-order asymptotic expansion using a score function. In this section, unlike their methods, we present the derivation of the fourth-order asymptotic UMVUE using the higher-order asymptotic expansion.

We define the general adjusted maximum likelihood estimator of \( \theta \) as follows:

\[
\hat{\theta}_{GA} = \arg\max_{\theta > 0} l_{ad}(\theta),
\]

where \( l_{ad}(\theta) = l(\theta) + \tilde{l}_{ad}(\theta) \).

In addition, we denote \( e^{l_{ad}} \) as the adjustment factor. For example, the maximum likelihood estimator \( \hat{\theta}_{ML} \) is obtained when \( \tilde{l}_{ad}(\theta) \propto C \) is adopted, where \( C \) is a constant value that does not depend on \( \theta \).

Next, we let \( \hat{R}_i^{(BCA)} \) be an estimator of \( R_i \), where \( \hat{R}_i \) in (5) is replaced with \( R_i(\hat{\theta}_{GA}) \). Theorem [3] is presented to show its property of asymptotic moments, of which the proof is shown in Appendix B.2.
Theorem 3. For large $n$ under regularity conditions $R1$ and $R2$,

\[(i) E[\tilde{R}_i^{(BCA)} - R_i(\theta)] = \frac{R_i}{g_\xi} \left[ \partial_\xi^{(1)} \tilde{l}_ad \left( 1 + \frac{\partial_\xi^{(2)} \tilde{l}_ad}{g_\xi} \right) - \frac{\partial_\xi^{(2)} \tilde{l}_ad \partial_\xi g_\xi}{2g_\xi^2} \right] + \frac{R_i}{2g_\xi^2} \left( g_\xi - \partial_\xi g_\xi - \partial_\xi^{(2)} \tilde{l}_ad + \partial_\xi^{(3)} \tilde{l}_ad \right) + o((\log n)^{-2}); \]

\[(ii) E[\{\tilde{R}_i^{(BCA)} - R_i(\theta)\}^2] = \frac{R_i^2}{g_\xi} \left( 1 + \frac{(\partial_\xi \tilde{l}_ad(\xi))^2 + 2\partial_\xi^{(2)} \tilde{l}_ad(\xi)}{g_\xi} \right) + o((\log n)^{-2}), \]

where $\xi = \log \theta$ and $g_\xi = \sum_{j=2}^{n} \frac{\theta^{(j-1)}}{(\theta+j-1)^2}$.

The theorem above implies that the following condition of the adjustment factor is required to eliminate the fourth-order asymptotic bias without sacrificing the asymptotic efficiency.

\[\partial_\xi \tilde{l}_ad(\xi) \left( 1 + \frac{\partial_\xi^{(2)} \tilde{l}_ad}{g_\xi} \right) - \frac{\partial_\xi^{(2)} \tilde{l}_ad \partial_\xi g_\xi}{2g_\xi^2} + \frac{1}{2g_\xi} \left( g_\xi - \partial_\xi g_\xi - \partial_\xi^{(2)} \tilde{l}_ad(\xi) + \partial_\xi^{(3)} \tilde{l}_ad(\xi) \right) = o((\log n)^{-1}), \quad (7)\]

\[(\partial_\xi \tilde{l}_ad(\xi))^2 + 2\partial_\xi^{(2)} \tilde{l}_ad(\xi) = o(1). \quad (8)\]

To obtain an adjustment factor satisfying (7) and (8), we therefore restrict the class of adjustment factor to the following for large $n$ with $j = 1, 2, 3$:

\[\partial_\theta^{(j)} \tilde{l}_ad(\theta) = o(1). \]

We then find that the resulting specific adjustment factor $e^{\tilde{l}_ad}$ satisfies

\[\partial_\theta \tilde{l}_ad(\theta) = -\frac{1}{\theta} g - \frac{\partial_\xi g}{2g} = -\frac{\sum_{j=2}^{n} \frac{(j-1)}{(\theta+j-1)^2}}{\sum_{j=2}^{n} \frac{(j-1)}{(\theta+j-1)^2}}. \quad (9)\]

Additionally, it holds that $\partial_\xi^{(j)} \tilde{l}_ad(\xi) = O((\log n)^{-1})$ with $j = 1, 2, 3$ from (16) given in Appendix A.

Subsequently, we let $\tilde{\theta}_A$ be the above-mentioned adjusted maximum likelihood estimator. It can be obtained as the root of the following equation:

\[\frac{K_2}{\theta} - \sum_{j=1}^{n} \frac{1}{\theta + j - 1} - \frac{\sum_{j=2}^{n} \frac{(j-1)}{(\theta+j-1)^2}}{\sum_{j=2}^{n} \frac{(j-1)}{(\theta+j-1)^2}} = 0. \]
In addition, we define the estimator \( \hat{R}^{(BC2)}_i \), which substitutes \( \hat{R}_i \) in (5) with \( R_i(\hat{\theta}_A) \). Corollary 1 summarizes the fourth-order asymptotic properties of its first and second moments.

**Corollary 1.** Under regularity condition R1 for large \( n \), the following holds:

\[
(i) \ E[\hat{R}^{(BC2)}_i - R_i(\theta)] = o((\log n)^{-2});
\]

\[
(ii) \ E[(\hat{R}^{(BC2)}_i - R_i(\theta))^2] = \frac{R_i^2}{\sum_{j=2}^{n} \theta_j (\theta_{j-1})^2} + o((\log n)^{-2}).
\]

Corollary 1 shows that our specific adjustment factor contributes to the disappearance of the fourth-order asymptotic bias without sacrificing asymptotic efficiency. As it is for \( \hat{R}^{(BC1)}_i \), the estimator is a function of the complete sufficient statistic \( K_n \) of \( \theta \) in cases \( \hat{\theta}_A \in \mathcal{S} \). Therefore, it is also an asymptotic UMVUE up to the fourth-order, based on Lemmas 1 and 2.

Next, we provide some remarks.

**Remark 3.1.** The estimator \( \hat{R}^{(BC1)}_i \) makes a downward correction by the term \( \hat{B}_i \) because the bias-corrected term \( \hat{B}_i \) is positive almost surely. Nevertheless, the term \( B_i \) has a slightly complex functional structure, which may result in unrealistic negative estimates of \( \hat{R}^{(BC1)}_i \). By contrast, \( \hat{R}^{(BC2)}_i \) maintains a simple function structure and ensures that it is in the range of \( R_i(\theta) \) simultaneously.

**Remark 3.2.** It is noteworthy that the logarithm of the adjustment factor \( \tilde{I}_{ad}(\theta) \) does not depend on \( i \) for obtaining \( \hat{\theta}_A \). In other words, \( \hat{\theta}_A \) can be used as one common estimate of \( \theta \) even in the simultaneous higher-order asymptotic inferences of \( R_i \) and \( \hat{R}_j \) (\( i \neq j \)), while maintaining the function form of \( R_i \). This may significantly reduce the computer burden in simultaneous inferences.

**Remark 3.3.** The bias correction for \( R_i(\theta) \) using our adjusted maximum likelihood method corresponds to the bias correction for \( \hat{\theta}_{ML} \), owing to the results (16) associated with the orders of \( \partial_\theta R_i \) and \( \partial_\theta^2 R_i \). One may recall that Firth (1993) also suggested a bias reduction method for \( \hat{\theta}_{ML} \). However, we expanded the higher-order asymptotic expansion and derived a fourth-order asymptotic UMVUE.

**Remark 3.4.** The Ewens model enables the asymptotic theoretical result of \( R_i \) to be constructed easily, owing to the properties of the exponential family distribution and the relationship \( \theta = e^\xi \), where \( \xi \) is the natural parameter.
For more details, see Appendix C.2. In general, a more complex proof may be required for the fourth-order asymptotic expansion.

**Remark 3.5.** Our adjustment factor obtained from (9) coincides with one example of Corollary 3.9 in Hirose and Mano (2021) for one-flat manifold, although the asymptotic orders of the Fisher information and expansion are different from those of our study.

4 Application to disclosure control: assessed risk of population unique

In official statistics, data providers often create secondary available tables from microdata for disclosure to users while guaranteeing security. In this case, the microdata are categorized by the attribution of individuals in the cells of the table. To protect personal information, the risk of individual identification must be assessed from such a table. This risk is referred to as the microdata disclosure risk.

In this example, $S_i$ represents the number of cells on the population, in which the number of individuals is $i$, whereas $K_N$ denotes the number of non-empty cells for population size $N$. Sibuya (1993) named the $S_i$ as \textit{size indices}.

The indices are used to assess the disclosure risk. For instance, when $S_i$ is large with a small $i$, it is interpretable that the table has a high disclosure risk. In practice, $S_i$ is estimated using $R_i$. In particular, $S_1$ is especially of interest to assess and is referred to as the number of \textit{population uniques}. By contrast, the cells in which individuals are unique in the sample is known as \textit{sample unique}. Additionally, the risk of “population and sample unique” is assessed through $f \times R_1$, where $f$ is a known sampling ratio. In such a case, $S_1$ should also be estimated because the number of sample uniques can be observed.

For the inference, super-population models are often used (Bethlehem et al., 1990; Hoshino and Takemura, 1998; Hoshino, 2001). In this study, we assume that the Ewens sampling formula is not only a super-population model, but also a sampling model. As mentioned earlier, this is reasonable because the Ewens sampling formula has a partition structure (Kingman, 1978).

It is clear that the previous methodology can be applied to estimate the number of population uniques as a disclosure risk. Theorem 1 realizes the...
asymptotic UMVUE of $R_1$, up to the order of $O((\log n)^{-1})$, as

$$\hat{R}_1^{(NM)} \equiv \begin{cases} 0 & (K_n = 1) \\ R_1(\hat{\theta}_{ML}) & (\hat{\theta}_{ML} \in \mathcal{S}) \\ R_1(C_+) & (\hat{\theta}_{ML} \notin \mathcal{S} \cup \{0\}) \end{cases},$$

where $C_+$ denotes a large but finite positive constant.

Next, from Theorems 2 and 3, $\hat{R}_1^{(BC1)}$ and $\hat{R}_1^{(BC2)}$ become the fourth-order asymptotic UMVUEs of $R_1$. Specifically, $\hat{R}_1^{(BC1)}$ is expressed as

$$\hat{R}_1^{(BC1)} \equiv \begin{cases} 0 & (K_n = 1) \\ R_1(\hat{\theta}_{ML}) \left[ 1 - \sum_{j=2}^{n} \frac{(j-1)}{[\hat{\theta}_{ML}+(j-1)]^2} \right] & (\hat{\theta}_{ML} \in \mathcal{S}) \\ R_1(C_+) & (\hat{\theta}_{ML} \notin \mathcal{S} \cup \{0\}) \end{cases},$$

An alternative estimator $\hat{R}_1^{(BC2)}$ provides a simpler formula, as follows:

$$\hat{R}_1^{(BC2)} \equiv \begin{cases} 0 & (K_n = 1) \\ R_1(\hat{\theta}_A) & (\hat{\theta}_A \in \mathcal{S}) \\ R_1(C_+) & (\hat{\theta}_A \notin \mathcal{S} \cup \{0\}) \end{cases},$$

provided that $\hat{\theta}_A > 0$.

As mentioned in Remark 2.1, the expectation parameter $\eta$ is used for the inference of the number of non-empty cells $K_N$ to obtain the exact UMVUE.

### 5 Monte-Carlo simulation

We implemented a finite sample simulation study to assess the efficiency of several estimators $\hat{R}_1$ through Monte-Carlo simulations.

Hence, we considered certain simulation settings such that population size $N = 10^4$, three sample size patterns, i.e., $n \in \{20, 10^2, 10^3\}$, and $10^4$ replications were generated from the Ewens distribution. Moreover, we set 15 (five values in each of the three patterns P1–P3) patterns of $\theta$ for each sample size $n$ to evaluate the relative effect of the true value of $\theta$ for sample size $n$, as follows: **P1**: $\theta \in \{1, 3, 5, 7, 9\}$; **P2**: $\theta \in \{10, 30, 50, 70, 90\}$; **P3**: $\theta \in \{100, 300, 500, 700, 900\}$.

Some cases existed where $n < \theta$, these asymptotic setting of which was not considered to obtain the theoretical result in this study. However, these results were also reported herein.
Three estimators of $R_1$ were considered for comparison, as follows: (i) the second-order asymptotic UMVUE $\hat{R}_1^{(NM)}$, introduced in Section 2; (ii) the fourth-order asymptotic UMVUE $\hat{R}_1^{(BC1)}$, introduced in Section 3.2; (iii) the fourth-order asymptotic UMVUE $\hat{R}_1^{(BC2)}$, introduced in Section 3.3. The estimators of (i)-(iii) are denoted as “NM,” “BC1,” and “BC2,” respectively. In addition, $C_\times = 10^6$ was adopted.

We first evaluated the estimators $\hat{R}_i$ using the relative bias and relative root of the mean squared error for the true $R_1$. The relative bias (RB) and relative root of the mean squared error (RRMSE) are defined as

$$RB \equiv \frac{1}{10^4 \times R_1} \sum_{r=1}^{10^4} (\hat{R}_1^{(r)} - R_1^{(r)}) \times 100,$$

$$RRMSE \equiv \frac{1}{R_1} \left\{ \frac{1}{10^4} \sum_{r=1}^{10^4} (\hat{R}_1^{(r)} - R_1^{(r)})^2 \right\}^{1/2} \times 100,$$

where an estimate $\hat{R}_1^{(r)}$ and a true value $R_1^{(r)}$ are constructed using the $r$th replication with $r = 1, \ldots, 10^4$.

Figure 1 shows the relative biases (RB) in nine figures for each combination of $(n, P)$, where $P$ denotes one of three patterns P1–P3 for $\theta$. The right side of the three figures show the results for case $n = 10^3$; as shown, all estimators demonstrated similar performance in terms of the relative bias. Meanwhile, in three other figures for cases $(n, P) \in \{(20, P1), (100, P1), (100, P2)\}$, two fourth-order asymptotic UMVUEs that outperformed the second-order asymptotic one are shown. In particular, in $(n, P) = (20, P1)$, $\hat{R}_1^{(BC2)}$ performed better than $\hat{R}_1^{(BC1)}$ in terms of the relative bias, as shown in the left of the top figures. Our asymptotic setting in this study did not consider the following simulation settings: $(n, P) \in \{(20, P2), (20, P3), (10^2, P3)\}$. Nonetheless, we also reported these results by changing the scale of the $y$ axis, although some results of $\hat{R}_1^{(BC1)}$ were not appeared because of their considerably low relative biases. Furthermore, these figures show that $\hat{R}_1^{(BC1)}$ can underestimate significantly, whereas the others performed similarly when $n$ is smaller than $\theta$. This might be caused by the inflation of $\hat{B}_1$, which may suggest another possibility for the theoretical differences between the fourth-order asymptotic UMVUEs in other asymptotic settings.

Next, the relative root of mean squared error (RRMSE) is shown in Figure 2, which comprises nine figures for each combination of $(n, P)$. The case $(n, P) = (20, P1)$ demonstrates the superiority of the fourth-order asymptotic UMVUEs in terms of the relative root of the mean squared error.
Figure 1: Relative biases (RBs) of three estimators of $R_1$ (NM: solid line; BC1: dashed line; and BC2: dotted line) in each combination $(n, P)$, where $P$ is one of three patterns P1–P3 for $\theta$, and sample size $n$ is fixed at 20 (left), $10^2$ (center), and $10^3$ (right); $x$ axis denotes values of $\theta$.

In cases $(n, P) \in \{(10^2, P1), (10^2, P2), (10^3, P1), (10^3, P2), (10^3, P3)\}$, we did not observe significant differences among all candidates from the figures. Moreover, as it is for Figure 1, we reported three other cases for $n < \theta$ with a scale change for the $y$ axis, and some results of $\hat{R}_1^{(BC1)}$ were not appeared because of their considerably large relative roots of the mean squared errors. Such results might be due to the considerable underestimation of $\hat{R}_1^{(BC1)}$. By contrast, no significant differences were observed between the other two estimators even in such cases.

Finally, we report the rate of occurrence of negative estimates of $R_1$ in Figure 3 for three cases: $(n, P) \in \{(20, P2), (20, P3), (100, P3)\}$, although we did not theoretically consider such cases in this study. We note that
negative estimates did not occur in other cases. The results showed that only the estimates of $\hat{R}_1^{(BC1)}$ can be negative, as shown in the three figures. In particular, the results with a relatively large $\theta$ for sample size $n$ imply a high probability of $\hat{R}_1^{(BC1)}$ being unrealistic negative estimates.

6 Conclusion and Discussion

In this study, we constructed three types of asymptotic UMVUE of $R_i$: one matched the moments of UMVUE up to the second-order, whereas the others, up to the fourth order. In addition, the non-existence of a solution
Figure 3: Rate of occurrence of negative estimates of three estimators of $R_1$ (NM: solid line; BC1: dashed line; and BC2: dotted line) in three combinations: $(n, P) \in \{(20, P_2), (20, P_3), (100, P_3)\}$; $x$ axis denotes values of $\theta$.

In parameter estimation, which is a serious practical problem, was avoided. Moreover, the common adjusted maximum likelihood estimator can be used in simultaneous inferences for $R_i$ and $R_j$ with $i \neq j$, similar to the maximum likelihood estimator. Furthermore, we applied this methodology to assess the risk of population unique in disclosure control.

In our case, the simulation study showed that the fourth-order UMVUE outperformed the second-order one in terms of the relative bias and relative root of the mean squared error. By contrast, in cases $n < \theta$, the higher-order asymptotic UMVUE using the additive bias correction method may result in an extremely low efficiency, as indicated from the simulation study. Moreover, such an estimator may pose a high risk of being negative estimates of $R_i$ in such cases.

These simulation results imply that some theoretical differences may occur among our asymptotic UMVUEs in other asymptotic settings: $n \to \infty$ and $\theta \to \infty$ as discussed in Tsukuda (2017). In the near future, we will attempt to address these issues.

Acknowledgements

The first author was supported in part by JSPS KAKENHI Grant 18K12758. The second author was supported in part by JSPS KAKENHI Grants 18H00835 and 20K03742.
Appendix A  Regularity conditions and related results

We give regularity conditions for several theorems in the following.

**R 1.** The parameter $\theta > 0$ and $i \in \mathbb{Z}_{>0}$ are bounded for large $n$. The sample size is $n \geq 2$.

**R 2.** The adjustment factor $\tilde{l}_ad(\xi)$ is in the class $C_6^\xi$, where $C_6^\xi$ is the set of the sixth-times differentiable functions of $\xi$ on $\mathbb{R}$. In addition, $\partial^{(j)}_\xi l_{ad}(\xi)$ does not depend on the random variable $K_n$ and are of the order $O(1)$ for $j = 1, \ldots, 6$.

Moreover, several calculation results are obtained, as follows:

\[
\partial_\theta R_i = \frac{\partial R_i(\xi)}{\partial \theta} = \frac{R_i}{\theta} \left[ 1 - \sum_{j=1}^{i} \frac{\theta}{\theta + N - j} \right],
\]

(10)

\[
\partial^2_\theta R_i(\theta) = \frac{R_i}{\theta} \left[ -2 \sum_{j=1}^{i} \frac{(N - j)}{\theta^2 + N - j)^2} + \sum_{j=1}^{i} \sum_{s \neq j} \frac{\theta}{\theta + N - j)(\theta + N - s)} \right].
\]

(11)

\[
g_\xi = g(\xi) = -\frac{\partial^2 l(\xi)}{\partial \xi^2} = \sum_{j=2}^{N} \frac{\theta(j - 1)}{(\theta + j - 1)^2};
\]

(12)

\[
\partial_\xi g = \sum_{j=2}^{N} \frac{\theta(j - 1)(j - 1 - \theta)}{(\theta + j - 1)^3}.
\]

(13)

\[
\partial^2_\xi g = \theta \partial_\theta g + 2 \theta^2 \partial^2_\theta g,
\]

\[
\partial^3_\xi g = \theta \partial_\theta g + 3 \theta^2 \partial^2_\theta g + \theta^3 \partial^3_\theta g,
\]

\[
\partial^4_\xi g = \theta \partial_\theta g + 7 \theta^2 \partial^2_\theta g + 6 \theta^3 \partial^3_\theta g + \theta^4 \partial^4_\theta g,
\]

(14)

where $\xi = \log \theta$ is a natural parameter of the exponential distribution family.

In addition, under regularity condition R1 with $j = 2, 3, 4$ and $l = 1, 2, 3, 4$, we have the following for large $n \leq N$:

\[
R_i \sim \frac{\theta}{i} = O(1), \quad \partial_\theta R_i \sim \frac{1}{i} = O(1), \quad \partial^2_\theta R_i = O(N^{-1}),
\]

\[
g_\xi \sim \theta \log n = O(\log n), \quad \partial_\xi g_\xi \sim \theta \log n = O(\log n), \quad \partial^2_\theta g_\xi = O(1),
\]

\[
g_\xi - \partial^2_\xi g = O(1), \quad B_i(\theta) = O((\log n)^{-2}), \quad \partial_\theta B_i(\theta) = O((\log n)^{-2}).
\]

(16)
To satisfy the first and second lines above, we use (10)–(13), whereas the second line and (12)–(15) are used for the last line.

Furthermore, we establish Lemma 3 to prove the theorems related to estimator \( \hat{\theta}_{GA} \) in (6). Subsequently, we redefine \( S = \{ \hat{\theta} \mid \hat{\theta} \in (0, C_+ + \] \} \).

**Lemma 3.** On set \( \mathcal{S} \), we have the following for large \( n \) under regularity conditions R1–R2:

\[
(i) E^\mathcal{S}[\hat{\theta}_{GA} - \theta] = \frac{\theta}{g_\xi} \left[ \partial_{\xi}^{(1)} \tilde{l}_{ad} \left( 1 + \frac{\partial_{\xi}^{(2)} \tilde{l}_{ad}}{g_\xi} \right) - \frac{\partial_{\xi}^{(2)} \tilde{l}_{ad} \partial_\xi g_\xi}{2g_\xi^2} \right] + \frac{\theta}{2g_\xi} \left( g_\xi - \partial_\xi g_\xi - \partial_{\xi}^{(2)} \tilde{l}_{ad} + \partial_{\xi}^{(3)} \tilde{l}_{ad} \right) + o((\log n)^{-2});
\]

\[
(ii) E^\mathcal{S}[\hat{\theta}_{GA} - \theta]^2 = \frac{\theta^2}{g_\xi} \left[ 1 + \frac{(\partial_\xi \tilde{l}_{ad})^2 + 2\partial_{\xi}^{(2)} \tilde{l}_{ad}(\xi)}{g_\xi} \right] + o((\log n)^{-2}),
\]

where \( E^\mathcal{S}[\cdot] \) is the expectation on set \( \mathcal{S} \).

The proof is shown in Appendix C.3.

**Appendix B**  **Proof of Theorems**

**Appendix B.1  Theorems [1–2]**

For Theorem 1, we use a method similar to that of Das et al. (2004) and Lemma 3.3 in Hirose and Mano (2021). Under the regularity condition, the following holds on \( \mathcal{S} \):

\[
E^\mathcal{S}[R_i(\hat{\theta}_{ML} - \theta) - R_i] = E^\mathcal{S}[\hat{\theta}_{ML} - \theta] \partial_\theta R_i + \frac{1}{2} E^\mathcal{S}[(\hat{\theta}_{ML} - \theta)^2] \partial_\theta^2 R_i \bigg|_{\theta = \theta^*},
\]

\[
= E^\mathcal{S}[\hat{\theta}_{ML} - \theta] \frac{R_i}{\theta} + o((\log n)^{-2}), \tag{17}
\]

where \( \theta^* \) lies between \( \theta \) and \( \hat{\theta}_{ML} \). In the above, the second equality holds from Lemma 3, (10), and (16).

From Lemma 3 (i), (10), (12), and (16), the following holds for large \( n \):

\[
\left. \frac{R_i}{\left( \sum_{j=2}^{n} \frac{(j-1)}{(\theta + j - 1)^3} \right)^2} \sum_{j=2}^{n} \frac{(j - 1)}{(\theta + j - 1)^3} + o((\log n)^{-2}) = O((\log n)^{-2}). \right)
\]

Hence, under regularity condition R1, we have the following for large \( n \):
\begin{align*}
|E[\hat{R}_i^{(NM)} - R_i]| & \leq |E^\mathcal{F} [\hat{R}_i^{(NM)} - R_i]| + |E^\mathcal{K} [\hat{R}_i^{(NM)} - R_i]| + |E^\mathcal{X}_i [\hat{R}_i^{(NM)} - R_i]|, \\
& \leq \frac{R_i}{\left\{ \sum_{j=2}^{n} \frac{\theta j (j-1)}{(\theta + j - 1)^2} \right\}^2} \sum_{j=2}^{n} \frac{(j - 1)}{(\theta + j - 1)^3} + C_R \{P(\mathcal{K}_c) + P(\mathcal{X}_1)\} + o((\log n)^{-2}), \\
& = \frac{R_i}{\left\{ \sum_{j=2}^{n} \frac{\theta j (j-1)}{(\theta + j - 1)^2} \right\}^2} \sum_{j=2}^{n} \frac{(j - 1)}{(\theta + j - 1)^3} + o((\log n)^{-2}), \quad (18)
\end{align*}

where $\mathcal{F}_K = \mathcal{F}^c \cap \mathcal{X}_1^c$ and $C_R = \sup |\hat{R}_i^{NM} - R_i|$. The $C_R$ is of the order $O(1)$ for large $n$ from the definition of the estimator of $R_i$. For the last equality to be valid, probabilities $P(\mathcal{F}_K)$ and $P(\mathcal{X}_1)$ are of the order $o((\log n)^{-2})$ for large $n$, as a result of Lemmas 1 and 2. Therefore, Theorem 1 (i) is shown from (18), along with Theorem 2 (i).

Next, we prove part (ii) for each theorem. From Lemma 3 (ii), (10), (11), and (16), we obtain the following for large $n$:

\begin{align*}
E^\mathcal{F} [(\hat{R}_i(\hat{\theta}_{ML}) - R_i)^2] & = E^\mathcal{F} [(\hat{\theta}_{ML} - \theta)^2](\partial_{\theta} R_i)^2 + o((\log n)^{-2}), \\
& = \frac{R_i^2}{\sum_{j=2}^{n} \frac{\theta (j-1)}{(\theta + j - 1)^2}} + o((\log n)^{-2}).
\end{align*}

Hence, Theorem 1 (ii) is obtained in a manner similar to (18), as follows:

\begin{align*}
E[(\hat{R}_i^{(NM)} - R_i)^2] = \frac{R_i^2}{\sum_{j=2}^{n} \frac{\theta (j-1)}{(\theta + j - 1)^2}} + o((\log n)^{-2}).
\end{align*}

For Theorem 2 (ii), it is expressed on $\mathcal{F}$ that

\begin{align*}
E^\mathcal{F} [(\hat{R}_i(\hat{\theta}_{ML}) - R_i - B_i(\hat{\theta}_{ML}))^2] & = E^\mathcal{F} [(\hat{R}_i(\hat{\theta}_{ML}) - R_i - (B_i(\hat{\theta}_{ML}) - B_i - B_i)^2], \\
& = \frac{R_i^2}{\sum_{j=2}^{n} \frac{\theta (j-1)}{(\theta + j - 1)^2}} + o((\log n)^{-2}).
\end{align*}

We used (16) for the second equality above. Hence, Theorem 2 (ii) is also shown in a similar manner to (18).
Appendix B.2 Theorem 3

For Theorem 3, we consider \( \hat{\theta}_{GA} \) introduced in Section 3.2. The following result is obtained under regularity conditions on set \( \mathcal{S} \):

\[
E_{\mathcal{S}} [R_i(\hat{\theta}_{GA}) - R_i] = E_{\mathcal{S}} [(\hat{\theta}_{GA} - \theta)R_i + \frac{1}{2} E_{\mathcal{S}} [(\hat{\theta}_{GA} - \theta)^2] \partial_\theta^2 R_i]_{\theta = \theta^*},
\]

where \( \theta^* \) lies between \( \theta \) and \( \hat{\theta}_{GA} \). In the equation above, we used Lemma 3, (10), and (16).

From Lemma 3 (i), it can be rewritten as follows for large \( n \):

\[
(19) = \sum_{j=2}^{n} \frac{\theta(j-1)}{\theta + 1} g_{j-1} g_j, \quad (19)
\]

where \( g_j = \sum_{j=2}^{n} \frac{\theta(j-1)}{\theta + 1} g_j \).

Hence, Theorem 3 is calculated similarly as (18).

Appendix C Proof of Lemmas

Appendix C.1 Lemma 1

Using (2), probabilities \( P(K_n = 1) \) and \( P(K_n = n) \) are calculated using the property of the unsigned Stirling number, as follows:

\[
P(K_n = 1) = \frac{\theta}{\theta^{[n]}} s(n, 1) = \frac{\theta \Gamma(\theta)(n - 1)!}{\Gamma(\theta + n)} \sim \Gamma(\theta + 1)n^{-\theta},
\]

\[
P(K_n = n) = \frac{\theta^n}{\theta^{[n]}} s(n, n) = \frac{\theta \Gamma(\theta)}{\sqrt{2\pi} n} e^{n\theta - n - \theta + 1/2}.\]
It is noted that Stirling’s formula was used in the calculations above. Hence,

\[ P(\mathcal{X}) = O(n^{-\theta} \vee e^{n\theta^\alpha}n^{-n-\theta+1/2}) = O(n^{-\theta}). \]

This lemma is then obtained from the equalities \( \lim_{n \to \infty} (\log n)^2 n^{-\theta} = 0. \)

**Appendix C.2 Lemma 2**

Theorem 2 in Das et al. (2004) provides \( P(\mathcal{B}^c) \sim O((\log n)^{-\nu/8}) \) with any \( \rho \in (0, 1) \) and a finite positive value \( \nu \). It is noted that \( \mathcal{B} \) is a set that satisfies for large \( n \), on \( \mathcal{B} \), \( \hat{\theta} \in \Theta, \partial_l(\hat{\theta}) = 0, |\hat{\theta} - \theta| < C_0(\log n)^{-\rho/2}, \) and

\[ \hat{\theta} = \theta + \frac{\partial_l l(\theta)}{g_\theta} + r_0, \]

where \( |r_0| \leq C_0(\log n)^{-\rho}u_* \) with \( E(u_*^\nu) \) being bounded and \( C_0 \) is a positive generic constant.

Let \( \mathcal{S}_1 \) be a set \( \{\hat{\theta} : |\hat{\theta} - \theta| < 2C_0\} \). Then it holds that \( \mathcal{S}_1 \subseteq \mathcal{S} \) under regularity condition R1 from the result \( \hat{\theta} > 0 \) when \( 1 < K_n < n \). Hence, we obtain \( \nu > 16 \),

\[ P(\mathcal{S}^c) \leq P(\mathcal{S}_1^c) \leq P(\mathcal{B}^c) = o((\log n)^{-2}). \]

In the above, the inequality \( P(\mathcal{B}) \leq P(\mathcal{S}_1^c) \) under regularity condition R1 is obtained from the result \( |\hat{\theta} - \theta| < C_0(\log n)^{-\rho/2} < 2C_0 \) for the case \( n \geq 2 \).

**Appendix C.3 Lemma 3**

We now establish a new lemma.

**Lemma 4.** Under regularity conditions R1 and R2, for \( j = 3, 4 \), it holds that

\[ E^\mathcal{S}[(\hat{\xi}_{GA} - \xi)^j] = O((\log n)^{-2}), \]

\[ E^\mathcal{S}[(\hat{\xi}_{GA} - \xi)^5] = O((\log n)^{-3}). \]

The proof is shown from (14)–(16) and Holder’s inequality. In addition, the following holds for the natural parameter \( \xi \) on the exponential family distribution:

\[ \mu_2 = \kappa_2, \mu_3 = \kappa_3, \mu_4 = \kappa_4 + 3\kappa_2^2, \mu_5 = \kappa_5 + 10\kappa_2\kappa_3, \mu_6 = \kappa_6 + 15\kappa_4\kappa_2 + 10\kappa_3^2 + 15\kappa_2^3, \]

\[ \kappa_j = \partial_{\xi}^{(j-2)} g_\xi, \quad (j = 2, \ldots, 6), \]

(20)
where $\mu_t$ and $\kappa_t$ are the central $t$-th moment and cumulant of $K_n$, given $n$.

In the model, we obtain the following using the natural parameter $\xi = \log \theta$:

$$E^\mathcal{S} [\hat{\theta}_{GA} - \theta] = E^\mathcal{S} \left[ \sum_{i=1}^{4} \frac{(\hat{\xi}_{GA} - \xi)^i}{i!} \right] e^\xi + E^\mathcal{S} \left[ \frac{(\hat{\xi}_{GA} - \xi)^5}{5!} \right] e^\xi^*, \tag{21}$$

where $\xi^*$ lies between $\xi$ and $\hat{\xi}_{GA}$ on $\mathcal{S}$ and $\hat{\xi}_{GA} = \log \hat{\theta}_{GA}$.

Meanwhile, under the regularity conditions, it holds that

$$-\partial l_{ad}(\xi) = \sum_{i=1}^{4} \frac{(\hat{\xi}_{GA} - \xi)^i}{i!} \partial l_{ad}(\xi) + \frac{(\hat{\xi}_{GA} - \xi)^5}{5!} \partial l_{ad}(\xi) \big|_{\xi = \xi^*},$$

$$= \sum_{i=1}^{4} \frac{(\hat{\xi}_{GA} - \xi)^i}{i!} \left[ \partial\left(\partial l_{ad}(\xi) - \{\partial l_{ad}(\xi) \right) \big|_{\xi = \xi^*}, \tag{22}$$

where $\xi^*$ lies between $\xi$ and $\hat{\xi}_{GA}$ on $\mathcal{S}$, and $r_1$ indicates the last terms on the right-hand side of the first line.

In Equation (22), it follows that $E^\mathcal{S} [r_1] = O((\log n)^{-2})$ on $\mathcal{S}$ from Lemma 4, (15), and (16).

Equation (22) can be rewritten as follows:

$$\sum_{i=1}^{4} \frac{(\hat{\xi}_{GA} - \xi)^i}{i!} = \frac{1}{g_\xi} \left[ \partial(\hat{\xi}_{GA} - \xi) l_{ad} + (\hat{\xi}_{GA} - \xi) \partial l_{ad}(\xi) - \frac{1}{2} (\hat{\xi}_{GA} - \xi)^2 \left( \partial l_{ad}^2 - \partial l_{ad}^3 \right) 

- \frac{1}{6} (\hat{\xi}_{GA} - \xi)^3 \left( \partial l_{ad}^4 - \partial l_{ad}^5 \right) - \frac{1}{4!} (\hat{\xi}_{GA} - \xi)^4 \left( \partial l_{ad}^6 - \partial l_{ad}^7 \right) \right] + r_2,$$

$$= \frac{1}{g_\xi} \left[ \partial l_{ad} + (\hat{\xi}_{GA} - \xi) \partial l_{ad} + \frac{1}{2g_\xi} \left(g_\xi - \partial g_\xi - \partial l_{ad}^2 + \partial l_{ad}^3 \right) \right] + r_3. \tag{23}$$

Note that Lemma 4, (15), and (16) are used to show $E^\mathcal{S} [r_2] = o((\log n)^{-2})$, where $r_2$ appears on the right side of the first equality. On the right side of the second equality, $r_3$ is expressed as

$$r_3 = -\frac{1}{g_\xi} \left[ \frac{1}{2} \left( (\hat{\xi}_{GA} - \xi)^2 - \frac{1}{g_\xi} \right) \left( \partial l_{ad}^2 - \partial l_{ad}^3 \right) 

+ \frac{1}{6} (\hat{\xi}_{GA} - \xi)^3 \left( \partial l_{ad}^4 - \partial l_{ad}^5 \right) + \frac{1}{4!} (\hat{\xi}_{GA} - \xi)^4 \left( \partial l_{ad}^6 - \partial l_{ad}^7 \right) \right] + r_2.$$
Moreover, the following holds for large $n$:

$$E^\mathcal{Y}[(\hat{\xi}_{GA} - \xi)^2] = \frac{1}{g_\xi} + o((\log n)^{-1}). \quad (24)$$

Subsequently, Lemma 4, (16), and (24) yield the following result: $E^\mathcal{Y}[r_3] = o((\log n)^{-2})$.

Hence, for large $n$, we obtain

$$E^\mathcal{Y}\left[\sum_{i=1}^{4} \frac{(\hat{\xi}_{GA} - \xi)^i}{i!}\right] = \frac{1}{g_\xi} \left[\partial^{(1)}_{\xi} \tilde{l}_{ad} \left(1 + \frac{\partial^{(2)}_{\xi} \tilde{l}_{ad}}{g_\xi}\right) - \frac{\partial^{(2)}_{\xi} \tilde{l}_{ad} \partial_\xi g_\xi}{2g^2_\xi}\right] + \frac{1}{2g^2_\xi} \left(g_\xi - \partial_\xi g_\xi - \partial^{(2)}_{\xi} \tilde{l}_{ad} + \partial^{(3)}_{\xi} \tilde{l}_{ad}\right) + o((\log n)^{-2}). \quad (25)$$

For the equality above to be valid, we used the following result, which is obtained from (22):

$$E^\mathcal{Y}[\hat{\xi}_{GA} - \xi] = \frac{1}{g_\xi} \left[\partial^{(1)}_{\xi} \tilde{l}_{ad} - \frac{\partial_\xi g_\xi}{2g_\xi}\right] + o((\log n)^{-1}).$$

Combining Lemma 4, (21), and (25), part (i) is obtained as follows:

$$E^\mathcal{Y}[\hat{\theta}_{GA} - \theta] = \frac{\theta}{g_\xi} \left[\partial^{(1)}_{\xi} \tilde{l}_{ad} \left(1 + \frac{\partial^{(2)}_{\xi} \tilde{l}_{ad}}{g_\xi}\right) - \frac{\partial^{(2)}_{\xi} \tilde{l}_{ad} \partial_\xi g_\xi}{2g^2_\xi}\right] + \frac{\theta}{2g^2_\xi} \left(g_\xi - \partial_\xi g_\xi - \partial^{(2)}_{\xi} \tilde{l}_{ad} + \partial^{(3)}_{\xi} \tilde{l}_{ad}\right) + o((\log n)^{-2}).$$

Next, we present part (ii). On set $\mathcal{S}$, using Lemma 4 similarly as (21), we obtain

$$E^\mathcal{Y}[\hat{\xi}_{GA} - \xi] = E^\mathcal{Y}\left[\left(\sum_{i=1}^{3} \frac{(\hat{\xi}_{GA} - \xi)^i}{i!}\right)^2\right] e^{2\xi} + o((\log n)^{-2}). \quad (26)$$

Furthermore, using (14), (23) and Lemma 4

$$\left(\sum_{i=1}^{3} \frac{(\hat{\xi}_{GA} - \xi)^i}{i!}\right)^2 = \frac{1}{g^2_\xi} \left[\left(\partial^{(1)}_{\xi} l_{ad}\right)^2 + 2(\hat{\xi}_{GA} - \xi)\partial^{(4)}_{\xi} l_{ad}\partial^{(2)}_{\xi} \tilde{l}_{ad}\right] + r_4,$n

$$= \frac{1}{g^2_\xi} \left[\left(\partial^{(1)}_{\xi} l(\xi) + \partial^{(1)}_{\xi} \tilde{l}_{ad}(\xi)\right)^2 + 2(\hat{\xi}_{GA} - \xi)\partial^{(1)}_{\xi} l_{ad}\partial^{(2)}_{\xi} \tilde{l}_{ad}\right] + r_4.$$
where $r_4$ satisfies $E[r_4] = o((\log n)^{-2})$.

Based on the result above, Equation (26) reduces to

$$(26) = \frac{\theta^2}{g_\xi} \left[ 1 + \frac{(\partial_\xi \tilde{l}_{ad}(\xi))^2 + 2\partial_\xi^{(2)} \tilde{l}_{ad}(\xi)}{g_\xi} \right] + o((\log n)^{-2}).$$

Hence, Lemma 3 is proved.

References

[1] Antoniak, C.E. (1974). Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *The Annals of Statistics*, 2, 1152-1174.

[2] Bethlehem, J. G., Keller, W. J., and Pannekoek, J. (1990). Disclosure control of microdata. *Journal of the American Statistical Association*, 85, 38-45.

[3] Crane, H. (2016). The ubiquitous Ewens sampling formula. *Statistical science*, 31, 1-19.

[4] Das, K., Jiang, J., and Rao, J. N. K. (2004). Mean squared error of empirical predictor. *The Annals of Statistics*, 32, 818-840.

[5] Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoretical population biology*, 3, 87-112.

[6] Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80, 27-38.

[7] Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, 40, 237-264.

[8] Hirose, M. Y., and Lahiri, P. (2018). Estimating variance of random effects to solve multiple problems simultaneously. *The Annals of Statistics*, 46, 1721-1741.

[9] Hirose, M. Y., and Mano, S. (2021) A Bayesian construction of asymptotically unbiased estimators. *arXiv* preprint.

[10] Hoshino, N. (2001). Applying Pitman’s sampling formula to microdata disclosure risk assessment. *Journal of Official Statistics*, 17(4), 499.
[11] Hoshino, N., and Takemura, A. (1998). Relationship between logarithmic series model and other superpopulation models useful for microdata disclosure risk assessment. *Journal of the Japan Statistical Society*, 28, 125-134.

[12] Kingman, J. F. (1978). The representation of partition structures. *Journal of the London Mathematical Society*, 2, 374-380.

[13] Lahiri, P., and Li, H. (2010). Generalized maximum likelihood method in linear mixed models with an application in small-area estimation. *In Tenth Islamic Countries Conference on Statistical Sciences Statistics for Development and Good Governance* (p. 158).

[14] Li, H. and Lahiri, P. (2010). An adjusted maximum likelihood method for solving small area estimation problems, *Journal of Multivariate Analysis*, 101, 882-892.

[15] Mano, S. (2018). *Partitions, hypergeometric systems, and Dirichlet processes in statistics*. Tokyo: Springer.

[16] Sibuya, M. (1993). A random clustering process. *Annals of the Institute of Statistical Mathematics*, 45, 459-465.

[17] Tavare, S., and Ewens, W. J. (1997). *The Multivariate Ewens distribution. Chapter 41 In: Johnson, N.L., Kotz, S., Balakrishnan, N. (eds.) Multivariate Discrete Distributions*. Wiley, New York.

[18] Tsukuda, K. (2017). Estimating the large mutation parameter of the Ewens sampling formula. *Journal of Applied Probability*, 42-54.

[19] Watterson, G. A. (1974). The sampling theory of selectively neutral alleles. *Advances in applied probability*, 463-488.

[20] Yoshimori, M. and Lahiri, P. (2014). A new adjusted maximum likelihood method for the Fay–Herriot small area model. *Journal of Multivariate Analysis* 124 281-294.