SEVERAL STATISTICAL RESULTS UNDER MULTINOMIAL DISTRIBUTION WITH INFINITE CATEGORIES

by

Jun Zhou

A dissertation submitted to the faculty of
the University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Applied Mathematics

Charlotte

2009

Approved by:

Dr. Zhiyi Zhang

Dr. Ming Dai

Dr. Yanqing Sun

Dr. Moutaz Khouja

ABSTRACT

JUN ZHOU. Several Statistical Results under Multinomial Distribution with Infinite Categories. (Under the direction of DR. ZHIYI ZHANG)

This dissertation discusses several statistical results under multinomial distribution with infinite categories. Firstly, the discussion focuses on Simpson's diversity index and Turing's formula. We established an unbiased estimate for the newly proposed Generalized Simpson's indices and the associated asymptotic properties and showed that the parameters of a multinomial distribution may be re-parameterized as a set of Generalized Simpson's diversity indices. Secondly, two-dimensional asymptotic normality of a non-parametric sample coverage estimate based on Turing's formulae was derived under a fixed underlying probability distribution $\{p_k; k = 1, 2, \cdots\}$ where all $p_k > 0$. Thirdly, the dissertation also establishes a previously unknown sufficient condition for the second order Turing's formula. The newly derived asymptotic results based on Turing's formula paves a possible way to establish a new estimating approach for Hill's tail probability model.

ACKNOWLEDGMENTS

Firstly, I would like to express my deepest appreciation to my wonderful advisor Dr. Zhiyi Zhang for his guidance, training and support during my research. I feel so fortunate to have worked with him and learned from him, and I have great admiration for his intellect and work ethic.

My thanks also go to Dr. Zongwu Cai, Dr. Yanqing Sun, Dr. Ming Dai, and Dr. Jiancheng Jiang for their kindly help with various problems in my study and research.

I also want to thank my dear friends, Dichao Peng, Wei Huang and Yi Shen for their support and encouragement during these years. They make me feel that life is so beautiful.

Finally I want to thank my family. The encouragement and support from my beloved wife Jiajia Sun is a powerful source of inspiration and energy. My special appreciation goes to my parents, for their selfless love and support!

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1: GENERALIZED SIMPSON'S DIVERSITY INDEX

1.1 Introduction

Simpson's diversity index is a measure of diversity. In ecology, it is often used to quantify the biodiversity of a habitat. It takes into account the number of species present, as well as the abundance of each species.

Consider a multinomial probability distribution with infinite categories indexed by a positive integer s, *i.e.*, $\{p_s\} = \{p_s;\ s = 1, 2, ...\}$ where p_s may be viewed as the proportion of s^{th} species in a population. Simpson (1949) defined a biodiversity index $\lambda = \sum_{s=1}^{S} p_s^2$ for a population with a finite number of species S, which has an equivalent form

$$\zeta_{1,1} = 1 - \lambda = \sum_{s=1}^{S} p_s q_s \tag{1}$$

where $q_s = 1 - p_s$. $\zeta_{1,1}$ assumes a value in $[0, 1)$ with a higher level of $\zeta_{1,1}$ indicating a more diverse population, and is widely used across many fields of study.

Simpson's biodiversity index can be naturally and beneficially generalized in two directions. First, the dimension of the underlying multinomial distribution may be extended to infinity. Second, $\zeta_{1,1}$ may be considered as a special member of the following family:

$$\zeta_{u,v} = \sum p_s^u q_s^v \tag{2}$$

where $u \geq 1$ and $v \geq 0$ are two arbitrarily fixed integers, $\sum = \sum_{s \geq 1}$ as will be observed in subsequent text of this chapter unless otherwise specified. (2) may be viewed as a weighted version of (1), *e.g.*, $\zeta_{1,2}$ loads higher weight on minor species (those with smaller p_s's), and $\zeta_{2,1}$ loads higher weight on major species (those with larger p_s's), etc.

In the literature of biodiversity, there exists a vast collection of indices. While all are designed to measure species richness in a population, these indices can roughly be classified into two main categories: 1) the unknown number of species S with non-zero probabilities

in the population; and 2) the distributional evenness of the species. The methodological discussions on indices in the first category seem to rely on various additional parametric structures of a prior distribution. Many important references can be found in Wang and Lindsay (2005) among others. One of the key elements of estimating indices of this type is the sample coverage which has many intriguing properties. Interested readers may refer to Good (1953) for an introduction, and Robbins (1968), Esty (1983), Zhang and Huang (2007), Zhang and Huang (2008) and Zhang and Zhang (2009) for its statistical properties. In the second category, many different diversity indices have been proposed. Among the most discussed are Simpson's index $\lambda = \sum p_s^2$, Shannon's index $\theta = -\sum p_s \ln(p_s)$, and the Rényi-Hill index $\mathcal{N}_\alpha = (\sum p_s^\alpha)^{1/(1-\alpha)}$ for $\alpha \geq 0$ proposed by Rényi (1961) and generalized by Hill (1973). All these indices are defined only for populations with finite number of species. There are a few functional relationships among these and other indices. For example, $\lambda = 1/\mathcal{N}_2$ and $\theta = \lim_{\alpha \to 1} \ln(\mathcal{N}_\alpha)$. For a comprehensive discussion on the various relationships among the indices, one may refer to Rennolls and Laumonier (2006). Among the three indices mentioned above, only Simpson's index may easily be extended to the case of populations with infinite number of species with guaranteed convergence under unrestricted $\{p_s\}$ while the series in the other two indices may diverge for some vector values of $\{p_s\}$.

However the focus on $\zeta_{u,v}$ in this paper is not only motivated by the fact that the generalization of Simpson's biodiversity index is natural both in extending the dimension of the underlying multinomial distribution from finite to infinite and in adopting weighting schemes on the population species. It is also motivated by the existence of a class of well-behaving estimators. While many diversity indices have been proposed in the ecological literature, surprisingly little is known about the associated estimators in terms of their statistical properties. The general approach to the estimation problem seems to be simply replacing the population proportions in the indices with the sample proportions $\hat{p}_s$. The nonlinearity of the functions seems to, not surprisingly, cause a common but serious problem in bias. Most of the proposed methodologies adopt some form of adjustment aiming at reducing the bias by various techniques. As a result, the adjusted estimators become more complex in form and their corresponding distributional characteristics become less tractable.

In most of the applications, techniques such as jackknife and bootstrap are the norm, for an example, see Fritsch and Hsu (1999). Even in the case of Simpson's index $\zeta_{1,1}$, no convincing asymptotic distributional characteristics were derived except in some naive approach (the replicate approach) in which the *iid* sample of size rn is arbitrarily split into r *iid* subsamples of size n. The asymptotic normality was then achieved by allowing n to increase to infinity. A description of the "replicate approach" may be found in Magurran (1988) or Rogers and Hsu (2001).

In the next section, it is shown that the two parameterizations, $\{p_s\}$ and $\{\zeta_{u,v}\}$, are equivalent up to a permutation on the index set $\{s\}$. In Section 3, for each fixed pair of integers $u \geq 1$ and $v \geq 0$, an unbiased estimator of $\zeta_{u,v}$ is proposed, and its asymptotic normality is established for all $\{p_s\}$ when $\{p_s\}$ contains infinitely many species with positive probabilities and for all non-uniform $\{p_s\}$ when $\{p_s\}$ contains finitely many species with positive probabilities. It is also established that in the special case of S being finite, known or unknown, the proposed estimator is uniformly minimum variance unbiased (umvu) for all $\{p_s\}$ and asymptotically efficient for all non-uniform $\{p_s\}$. In Section 4, results of several simulation studies are reported to assess the adequacy of the asymptotic normality for various sample size n.

1.2 Re-parameterization

Let $\mathbf{P}$ be the parameter space where $\{p_s\}$ resides. Let O be a mapping that maps each $\{p_s\} \in \mathbf{P} \subset R^\infty$ to a non-increasingly ordered array $\{p_s\} \in R^\infty$. Let $\mathbf{P'} = O(\mathbf{P})$. For each $\{p_s\} \in \mathbf{P'}$ and each positive integer $u \geq 1$, let $\zeta_u = \zeta_u(\{p_s\}) = \sum p_s^u$ and $\{\zeta_u\} = \{\zeta_u; u \geq 1\}$. Consider the mapping from $\mathbf{P'}$ to $\mathbf{Z'} = M(\mathbf{P'}) \subset R^\infty$:

$$M : \{p_s\} \to \{\zeta_u\}. \tag{3}$$

Theorem 1. M in (3) is injective.

Proof. For every $\{p_s\} \in \mathbf{P'}$, $M(\{p_s\})$ is unique. It suffices to show that, for every $\{\zeta_u\} \in \mathbf{Z'}$, $M^{-1}(\{\zeta_u\})$ is unique. Suppose that there existed two sequences, $\{p_s\}$ and $\{q_s\}$, in $\mathbf{P'}$

satisfying $\sum p_s^u = \sum q_s^u$ for all $u \geq 1$. Let $s_0 = \min\{s; p_s \neq q_s\}$. If s_0 does not exist, then $\{p_s\} = \{q_s\}$. If s_0 existed, then

$$\sum_{s \geq s_0} p_s^u = \sum_{s \geq s_0} q_s^u \tag{4}$$

for all $u \geq 1$. It can be easily shown that

$$1 \leq r_p = \lim_{u \to \infty} \frac{\sum_{s \geq s_0} p_s^u}{p_{s_0}^u} < \infty \qquad and \qquad 1 \leq r_q = \lim_{u \to \infty} \frac{\sum_{s \geq s_0} q_s^u}{q_{s_0}^u} < \infty \tag{5}$$

where r_p and r_q are multiplicities of p_s's with the same value as p_{s_0} and of q_s's with the same value as q_{s_0} respectively. But by (4),

$$\frac{\sum_{s \geq s_0} p_s^u}{p_{s_0}^u} = \frac{\sum_{s \geq s_0} q_s^u}{q_{s_0}^u} \left(\frac{q_{s_0}}{p_{s_0}} \right)^u. \tag{6}$$

the right side of (6) approaches 0 or ∞ as $u \to \infty$ if $p_{s_0} \neq q_{s_0}$, which contradicts (5). Therefore s_0 does not exist and $\{p_s\} = \{q_s\}$.

It is to be noted that the monotonicity condition on $\{p_s\}$ cannot be further relaxed. This is because $\{\zeta_u\}$ is invariant under any permutation of the index set $\{s\}$ and $\{p_s\}$ is not. The one-to-one correspondence between $\mathbf{P}'$ and $\mathbf{Z}'$ via M is and can only be established under the monotonicity condition.

Theorem 1 has an intriguing implication: the complete knowledge of $\{p_s\}$ up to a permutation and the complete knowledge of $\{\zeta_u\}$ are equivalent. On the other hand, letting $\mathbf{Z} = \{\zeta_{u,v}; u \geq 1, v > 0\}$, each member of $\mathbf{Z}$ is a linear combination of finite members of $\mathbf{Z}'$. Therefore the complete knowledge of $\{p_s\}$ up to a permutation and the complete knowledge of $\{\zeta_{u,v}\}$ are equivalent. In other words, all the Generalized Simpson's diversity indices collectively and uniquely determine the underlying distribution. This implication is another motivation for Generalizing Simpson's diversity index beyond $\zeta_{1,1}$.

1.3 Estimators

Let X_i, $i = 1, \cdots, n$ be an *iid* sample under $\{p_s\}$. X_i may be written as $X_i = (X_{i,s}; s \geq 1)$ where for every i, $X_{i,s}$ takes 1 only for one s and 0 for all other s values. Let $Y_s =$

$\sum_{i=1}^{n} X_{i,s}$ and $\hat{p}_s = Y_s/n$. Y_s is the number of observations of the s^{th} species found in the sample. The following is the proposed estimator for $\zeta_{u,v}$.

$$Z_{u,v} = \binom{n}{u+v}^{-1} \binom{u+v}{u}^{-1} \sum_{s \geq 1} \left[1_{[Y_s \geq u]} \binom{Y_s}{u} \binom{n-Y_s}{v} \right]. \tag{7}$$

$Z_{u,v}$ is a function of $\{Y_s; s \geq 1\}$ and hence of $\{\hat{p}_s\} = \{\hat{p}_s; s \geq 1\}$. For a few special pairs of u and v, $Z_{u,v}$ reduces to

$$\begin{aligned}
Z_{1,1} &= \tfrac{n}{n-1} \textstyle\sum_{[\hat{p}_s \geq 1/n]} \hat{p}_s(1-\hat{p}_s) \\
Z_{2,0} &= \tfrac{n}{n-1} \textstyle\sum 1_{[\hat{p}_s \geq 2/n]} \hat{p}_s(\hat{p}_s - 1/n) \\
Z_{3,0} &= \tfrac{n^2}{(n-1)(n-2)} \textstyle\sum 1_{[\hat{p}_s \geq 3/n]} \hat{p}_s(\hat{p}_s - 1/n)(\hat{p}_s - 2/n) \\
Z_{2,1} &= \tfrac{n^2}{(n-1)(n-2)} \textstyle\sum 1_{[\hat{p}_s \geq 2/n]} \hat{p}_s(\hat{p}_s - 1/n)(1-\hat{p}_s) \\
Z_{1,2} &= \tfrac{n^2}{(n-1)(n-2)} \textstyle\sum 1_{[\hat{p}_s \geq 1/n]} \hat{p}_s(1-\hat{p}_s)(1-1/n-\hat{p}_s).
\end{aligned} \tag{8}$$

$Z_{u,v}$ is an unbiased estimator of $\zeta_{u,v}$. This fact is established by a U-statistic construction of the estimator. Let $m = u + v$. For every sub-sample of size m, say $\{X_1, \cdots, X_m\}$, consider the number of species in the population that are represented exactly u times in the sub-sample, *i.e.*, $N_u = \sum 1_{\left[\sum_{i=1}^{m} X_{i,s} = u\right]}$.

$$E(N_u) = \sum P\left[\sum_{i=1}^{m} X_{i,s} = u\right] = \sum \binom{m}{u} p_s^u q_s^v.$$

Therefore $\binom{u+v}{u}^{-1} N_u$ is an unbiased estimator of $\zeta_{u,v}$. There are a total of $K = \binom{n}{m}$ distinct sub-samples of size m, and therefore

$$\tilde{Z}_{u,v} = \binom{n}{u+v}^{-1} \binom{u+v}{u}^{-1} \sum_{k=1}^{K} N_u^{(k)}$$

where k indexes a particular sub-sample is an unbiased estimator of $\zeta_{u,v}$. On the other hand, $\sum_{k=1}^{K} N_u^{(k)}$ is simply the total number of times exactly u observations are found in a same species among all possible sub-samples of size m taken from the sample of size n. In counting the total number of such events, it is to be noted that, for a fixed u, only for species that are represented in the sample u times or more can such an event occur. Therefore $\sum_{k=1}^{K} N_u^{(k)} = \sum_{s\geq 1} 1_{[Y_s \geq u]} \binom{Y_s}{u}\binom{n-Y_s}{v}$ and hence $Z_{u,v} \equiv \tilde{Z}_{u,v}$.

The above U-statistic construction paves the path for establishing the asymptotic normality of $Z_{u,v}$. Let $X_1, \cdots, X_n$ be an *iid* sample under a distribution F, $\theta = \theta(F)$ be a parameter of interest, $h(X_1, \cdots, X_m)$ where $m < n$ be a symmetric kernel satisfying $E_F\{h(X_1, \cdots, X_m)\} = \theta(F)$, $U_n = U(X_1, \cdots, X_n) = \binom{n}{m}^{-1} \sum_k h(X_{1_k}, \cdots, X_{m_k})$ where the summation $\sum_k$ is over all possible sub-samples of size m from the sample of size n, $h_1(x_1) = E_F\{h(x_1, X_2, \cdots, X_m)\}$ be the conditional expectation of h given $X_1 = x_1$, and $\sigma_1^2 = Var_F\{h_1(X_1)\}$. The following lemma is by Hoeffding (1948).

Lemma 1. If $E_F\{h^2\} < \infty$ and $\sigma_1^2 > 0$, then $\sqrt{n}(U_n - \theta) \xrightarrow{d} N(0, m^2\sigma_1^2)$.

Let $C_k^r = k!/[r!(k-r)!]$ for any two non-negative integers k and r satisfying $k \geq r$. Let $m = u + v$ and $h = h(X_1, \cdots, X_m) = (C_m^u)^{-1} N_u$. Let $\mathbf{p} = \{p_s\}$. Suppose $u \geq 1$ and $v \geq 1$. Given $X_1 = x_1$,

$$
\begin{aligned}
C_m^u h_1(x_1) &= C_m^u E_{\mathbf{P}}\{h(x_1, X_2, \cdots, X_m)\} = E_{\mathbf{P}}\{N_u | X_1 = x_1\} \\
&= \sum 1_{[x_{1s}=1]} C_{m-1}^{u-1} p_s^{u-1} q_s^v + \sum 1_{[x_{1s}=0]} C_{m-1}^u p_s^u q_s^{v-1} \\
&= \sum C_{m-1}^u p_s^u q_s^{v-1} + \sum 1_{[x_{1s}=1]} C_{m-1}^u p_s^{u-1} q_s^{v-1} \left(q_s \tfrac{u}{v} - p_s\right) \\
&= C_{m-1}^u \sum p_s^u q_s^{v-1} + C_{m-1}^u \sum 1_{[x_{1s}=1]} p_s^{u-1} q_s^{v-1} \left(q_s \tfrac{u}{v} - p_s\right).
\end{aligned}
$$

$$
\begin{aligned}
(C_m^u)^2\sigma_1^2(u,v) &= (C_m^u)^2 Var_{\mathbf{P}}\{h_1(X_1)\} = \left(C_{m-1}^u\right)^2 Var_{\mathbf{P}}\left\{\sum 1_{[x_{1s}=1]} p_s^{u-1} q_s^{v-1}\left(q_s\tfrac{u}{v} - p_s\right)\right\} \\
&= \left(C_{m-1}^u\right)^2\left\{E_{\mathbf{P}}\left[\sum 1_{[x_{1s}=1]} p_s^{u-1} q_s^{v-1}\left(q_s\tfrac{u}{v} - p_s\right)\right]^2 - \left[\sum p_s^u q_s^{v-1}\left(q_s\tfrac{u}{v} - p_s\right)\right]^2\right\} \\
&= \left(C_{m-1}^u\right)^2\left\{\sum p_s^{2u-1} q_s^{2v-2}\left(q_s\tfrac{u}{v} - p_s\right)^2 - \left[\sum p_s^u q_s^{v-1}\left(q_s\tfrac{u}{v} - p_s\right)\right]^2\right\} \\
&= \tfrac{u^2}{v^2}\left(C_{m-1}^u\right)^2\sum p_s^{2u-1} q_s^{2v} - \tfrac{2u}{v}\left(C_{m-1}^u\right)^2\sum p_s^{2u} q_s^{2v-1} + \left(C_{m-1}^u\right)^2\sum p_s^{2u+1} q_s^{2v-2} \\
&\quad - \left(C_{m-1}^u\right)^2\left(\tfrac{u}{v}\sum p_s^u q_s^v - \sum p_s^{u+1} q_s^{v-1}\right)^2 \\
&= \tfrac{u^2}{v^2}\left(C_{u+v-1}^u\right)^2\zeta_{2u-1,2v} - \tfrac{2u}{v}\left(C_{u+v-1}^u\right)^2\zeta_{2u,2v-1} + \left(C_{u+v-1}^u\right)^2\zeta_{2u+1,2v-2} \\
&\quad - \left(C_{u+v-1}^u\right)^2\left(\tfrac{u}{v}\zeta_{u,v} - \zeta_{u+1,v-1}\right)^2 \geq 0.
\end{aligned} \tag{9}
$$

The last inequality in (9) becomes an equality only when $h(X_1)$ is a constant which occurs only when all the positive probabilities of $\{p_s\}$ are equal. Furthermore, since N_u is bounded for every fixed m, $E_{\{p_s\}}\{h^2\} < \infty$ is obviously true.

The following definition helps to simplify the subsequent presentation.

Definition 1. A multinomial distribution $\{p_s\} = \{p_s; s \geq 1\}$ is said to be uniform if all the non-zero probabilities of $\{p_s\}$ are identical.

Definition 1 implies that $\{p_s\}$ must not be a uniform distribution if it has infinitely many non-zero probabilities.

Suppose $u \geq 1$ and $v = 0$, therefore $C_m^u = 1$. It is easy to see that $h_1(x_1) = \sum 1_{[x_{1s}=1]} p_s^{u-1}$ and

$$
\sigma_1^2(u,0) = Var_{\mathbf{P}}\{h_1(X_1)\} = \sum p_s^{2u-1} - \left(\sum p_s^u\right)^2 = \zeta_{2u-1,0} - \zeta_{u,0}^2 \geq 0. \tag{10}
$$

The strict inequality holds for all cases except when $\{p_s\}$ is uniform.

Thus the following theorem is established.

Theorem 2. If $\{p_s\}$ is a non-uniform multinomial distribution, then for any given pair of positive integers u and v, $Z_{u,v}$ in (7), $\zeta_{u,v}$ in (2), $\sigma_1^2(u,v)$ in (9), and $\sigma_1^2(u,0)$ in (10),

$$\sqrt{n}(Z_{u,v}-\zeta_{u,v}) \xrightarrow{d} N(0,(u+v)^2\sigma_1^2(u,v)) \quad \text{and} \quad \sqrt{n}(Z_{u,0}-\zeta_{u,0}) \xrightarrow{d} N(0,u^2\sigma_1^2(u,0)). \tag{11}$$

Theorem 2 immediately implies consistency of $Z_{u,v}$ of $\zeta_{u,v}$ and the consistency of $Z_{u,0}$ of $\zeta_{u,0}$ for any $u \geq 1$ and $v \geq 1$ under the stated condition.

By the last expression of (9), (10) and Theorem 2, it is easily seen that when $u \geq 1$ and $v \geq 1$,

$$\begin{aligned} \hat{\sigma}_1^2(u,v) &= \left(\tfrac{v}{u+v}\right)^2 \left[\tfrac{u^2}{v^2} Z_{2u-1,2v} - \tfrac{2u}{v} Z_{2u,2v-1} + Z_{2u+1,2v-2} - \left(\tfrac{u}{v} Z_{u,v} - Z_{u+1,v-1}\right)^2\right], \text{ and} \\ \hat{\sigma}_1^2(u,0) &= Z_{2u-1,0} - Z_{u,0}^2 \end{aligned} \tag{12}$$

are consistent estimators of $\sigma_1^2(u,v)$ and of $\sigma_1^2(u,0)$ respectively, and hence the following corollary is established.

Corollary 1. If the condition of Theorem 2 is satisfied, then for any given pair of positive integers u and v, $Z_{u,v}$ in (7), $\zeta_{u,v}$ in (2), $\hat{\sigma}_1^2(u,v)$ and $\hat{\sigma}_1^2(u,0)$ in (12),

$$\frac{\sqrt{n}(Z_{u,v}-\zeta_{u,v})}{(u+v)\hat{\sigma}_1(u,v)} \xrightarrow{d} N(0,1) \quad \text{and} \quad \frac{\sqrt{n}(Z_{u,0}-\zeta_{u,0})}{u\hat{\sigma}_1(u,0)} \xrightarrow{d} N(0,1). \tag{13}$$

As a case of special interest when $u = v = 1$, the computational formula of $Z_{1,1}$ is given in (8) and

$$\frac{\sqrt{n}(Z_{1,1}-\zeta_{1,1})}{2\hat{\sigma}_1(1,1)} \xrightarrow{d} N(0,1) \tag{14}$$

where $\hat{\sigma}_1(1,1)$ is such that $4\hat{\sigma}_1^2(1,1) = Z_{1,2} - 2Z_{2,1} + Z_{3,0} - (Z_{1,1} - Z_{2,0})^2$ and $Z_{1,2}$, $Z_{2,1}$, $Z_{3,0}$ and $Z_{2,0}$ are all given in (8). (14) may be used for large sample inferences with respect to Simpson's index, $\zeta_{1,1}$, whenever the non-uniformity of the underlying multinomial distribution is considered as reasonable.

$Z_{u,v}$ is an umvue of $\zeta_{u,v}$ when S is finite. Since $Z_{u,v}$ is unbiased, by the Lehmann-Scheffe Theorem it suffices to show that $\{\hat{p}_s\}$ is a set of complete and sufficient statistics under $\{p_s\}$. When S is finite and known, under the multinomial assumption, $\{\hat{p}_s\}$ is complete and sufficient. When S is finite but unknown, $\{\hat{p}_s\}$ is obviously sufficient. The completeness is established by the following argument: by the definition of complete statistics, it is to be shown that for any function $g(\{\hat{p}_s\})$ satisfying $E[g(\{\hat{p}_s\})] = 0$ for each $(S, \{p_s\})$ implies $P\{g(\{\hat{p}_s\}) = 0\} = 1$ for each $(S, \{p_s\})$. If $E[g(\{\hat{p}_s\})] = 0$ for each $(S, \{p_s\})$ then for each fixed S, $E[g(\{\hat{p}_s\})] = 0$ for each $\{p_s\}$ since $\{\hat{p}_s\}$ is complete for the multinomial distribution, it follows that $P[g(\{\hat{p}_s\}) = 0] = 1$ for each $\{p_s\}$. Now S is arbitrary, thus one actually has $E[g(\{\hat{p}_s\})] = 0$ for each $(S, \{p_s\})$ implies $P\{g(\{\hat{p}_s\}) = 0\} = 1$ for each $(S, \{p_s\})$.

$Z_{u,v}$ is asymptotically efficient when S is finite. This fact is established by recognizing first that $\{\hat{p}_s\}$ is the maximum likelihood estimator (mle) of $\{p_s\}$, second that $\hat{\zeta}_{u,v} = \sum \hat{p}_s^u (1 - \hat{p}_s)^v$ is the mle of $\zeta_{u,v}$, and third that $\sqrt{n}(Z_{u,v} - \hat{\zeta}_{u,v}) \to 0$ in probability. To see the third fact, consider the following expression of $Z_{u,v}$ which may be obtained by a few algebraic manipulations from (7).

$$Z_{u,v} = \frac{n^{u+v}[n-(u+v)]!}{n!} \sum_{s=1}^{S} \left\{ 1_{[\hat{p}_s \geq u/n]} \prod_{i=0}^{u-1} \left(\hat{p}_s - \frac{i}{n} \right) \left[1_{[v=0]} + 1_{[v \geq 1]} \prod_{j=0}^{v-1} \left(1 - \hat{p}_s - \frac{j}{n} \right) \right] \right\}. \tag{15}$$

Since the coefficient in front of the summation in (15) converges to 1 as $n \to \infty$, it is only to show that

$$\sqrt{n} \left\{ \sum_{s=1}^{S} \left\{ 1_{[\hat{p}_s \geq u/n]} \prod_{i=0}^{u-1} \left(\hat{p}_s - \frac{i}{n} \right) \left[1_{[v=0]} + 1_{[v \geq 1]} \prod_{j=0}^{v-1} \left(1 - \hat{p}_s - \frac{j}{n} \right) \right] \right\} - \hat{\zeta}_{u,v} \right\} \xrightarrow{p} 0,$$

or letting $\hat{\zeta}_{u,v} = \sum 1_{[\hat{p}_s \geq u/n]} \hat{p}_s^u (1-\hat{p}_s)^v + \sum 1_{[\hat{p}_s < u/n]} \hat{p}_s^u (1-\hat{p}_s)^v \quad (\stackrel{def}{=} \hat{\zeta}_{u,v}^{(1)} + \hat{\zeta}_{u,v}^{(2)})$,

$$\sqrt{n}\left\{\sum_{s=1}^{S}\left\{1_{[\hat{p}_s \geq u/n]}\prod_{i=0}^{u-1}\left(\hat{p}_s - \frac{i}{n}\right)\left[1_{[v=0]} + 1_{[v\geq 1]}\prod_{j=0}^{v-1}\left(1-\hat{p}_s-\frac{j}{n}\right)\right]\right\} - \hat{\zeta}_{u,v}^{(1)}\right\} - \sqrt{n}\hat{\zeta}_{u,v}^{(2)} \xrightarrow{p} 0. \tag{16}$$

It is to show that each of the two terms in (16) converges to zero in probability.

First consider the case of $v = 0$. $\prod_{i=0}^{u-1}\left(\hat{p}_s - \frac{i}{n}\right)$ may be written as a sum of $\hat{p}_s^u$ and finitely many other terms each of which has the following form:

$$\frac{k_1}{n^{k_2}}\hat{p}_s^{k_3}$$

where k_1, $k_2 \geq 1$ and $k_3 \geq 1$ are finite fixed integers. Since

$$0 \leq \sqrt{n}\sum_{s=1}^{S} 1_{[\hat{p}_s \geq u/n]}\frac{|k_1|}{n^{k_2}}\hat{p}_s^{k_3} \leq \sqrt{n}\sum_{s=1}^{S} 1_{[\hat{p}_s \geq u/n]}\frac{|k_1|}{n^{k_2}}\hat{p}_s < \sqrt{n}\frac{|k_1|}{n^{k_2}} \to 0 \qquad \text{as } n \to \infty,$$

the first term of (16) converges to zero in probability. The second terms of (16) converges to zero when $u = 1$ is an obvious case since $\hat{\zeta}_{u,v}^{(2)} = 0$. It also converges to zero in probability when $u \geq 2$ since there are at most n terms in the sum and

$$0 \leq \sqrt{n}\sum_{s=1}^{S} 1_{[\hat{p}_s < u/n]}\hat{p}_s^u \leq \sqrt{n}\sum_{s=1}^{S} 1_{[\hat{p}_s < u/n]}[(u-1)/n]^u \leq (u-1)^u\sqrt{n}n/n^u \to 0.$$

Next consider the case of $v \geq 1$. $\prod_{i=0}^{u-1}\left(\hat{p}_s - \frac{i}{n}\right)\prod_{j=0}^{v-1}\left(1-\hat{p}_s-\frac{j}{n}\right)$ may be written as a sum of $\hat{p}_s^u(1-\hat{p}_s)^v$ and finitely many other terms each of which has the following form:

$$\frac{k_1}{n^{k_2}}\hat{p}_s^{k_3}(1-\hat{p}_s)^{k_4}$$

where k_1, $k_2 \geq 1$, $k_3 \geq 1$, and $k_4 \geq 1$ are finite fixed integers. Since

$$0 \leq \sqrt{n}\sum_{s=1}^{S} 1_{[\hat{p}_s \geq u/n]}\frac{|k_1|}{n^{k_2}}\hat{p}_s^{k_3}(1-\hat{p}_s)^{k_4} \leq \sqrt{n}\sum_{s=1}^{S} 1_{[\hat{p}_s \geq u/n]}\frac{|k_1|}{n^{k_2}}\hat{p}_s < \sqrt{n}\frac{|k_1|}{n^{k_2}} \to 0 \qquad \text{as } n \to \infty,$$

the first term of (16) converges to zero in probability. The second term of (16) converges to zero when $u = 1$ is an obvious case since $\hat{\zeta}_{u,v}^{(2)} = 0$. It also converges to zero in probability

when $u \geq 2$ since there are at most n terms in the sum and

$$0 \leq \sqrt{n}\sum_{s=1}^{S} 1_{[\hat{p}_s < u/n]}\hat{p}_s^u(1-\hat{p}_s)^v \leq \sqrt{n}\sum_{s=1}^{S} 1_{[\hat{p}_s < u/n]}\hat{p}_s^u \leq (u-1)^u n^{3/2}/n^u \to 0.$$

Thus the asymptotic efficiency of $Z_{u,v}$ is established.

1.4 Simulation Results

Twelve cases of simulation studies, four distributions by three levels of sample size, are conducted to examine the adequacy of the normal approximation in (14). The distributions used in the simulations studies are:

a. Triangular with $p_s = 0.02(s - 0.5)$, $s = 1, \cdots, 10$.

a. Finite Exponential with $p_s = ce^{-s/3}/3$, $s = 1, \cdots, 10$, where $c = (\sum_{s=1}^{10} e^{-s/3}/3)^{-1}$.

b. Pareto with $p_1 = p_2 = 1/3$, and $p_s = 2/[4(s-1)^2 - 1]$ for $s \geq 3$.

c. Exponential with $p_s = e^{-\frac{s-1}{10}} - e^{-\frac{s}{10}}$ for $s \geq 1$.

Each distribution is crossed with three levels of sample size, $n = 100$, $n = 500$ and $n = 1000$. Each simulation study is based on 1000 replications. Q-Q plots against $N(0,1)$ are given in Figure 1, with each row corresponding to a distribution in the order of the list above. The horizontal axis in each of the Q-Q plots is $N(0,1)$ and the vertical axis is the left-hand side of (14). The range on each axis is from -3 to 3. Columns 1, 2 and 3 in Figure 1.1 are corresponding to sample size levels 100, 500 and 1000 respectively.

Figure 1.1 indicates that the normality approximation of (14) is satisfactory within the range of -3 to 3 when $n = 500$ and $n = 1000$. For the cases of $n = 100$, only in the Pareto case which has a long thick right tail, the normality approximation is satisfactory. In the other three cases, which all have short (either finite or very thin right tail) tails, the sampling distributions of the left-hand side of (14) all seem to have thicker right tails than the standard normal distribution.

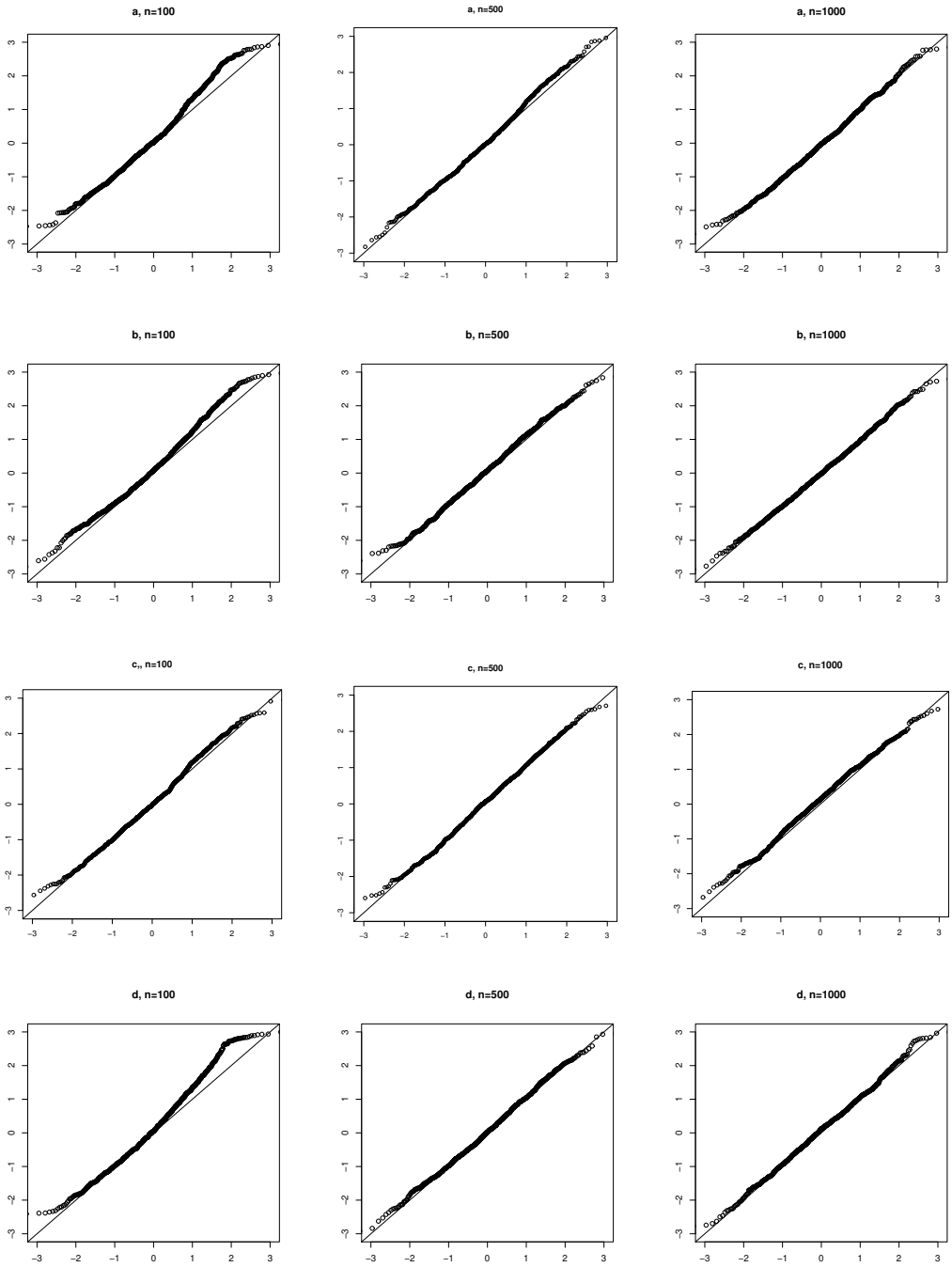

Figure 1.1: Q-Q plots for simulated data

1.5 Some Comments

The use of diversity indices is common but is not without doubts. One usual is that a single index cannot effectively capture the diversity of a population. Such a statement is

valid but is not a discredit to a particular index. The concept of diversity is not precisely defined and therefore no index could possibly be expected to capture the somewhat arbitrarily and often subjectively perceived diversity. On this front, the class of Generalized Simpson's indices proposed in this chapter offers a panel of estimable indices, which could potentially capture a wider range of diversity.

For (7) to be unbiased, $m = u + v$ must be less or equal to the sample size n. However for (13) to hold, $m = u + v$ must satisfy $2u + 2v - 1 \leq n$ or $u + v \leq (n + 1)/2$. This is indeed a restriction on the choices of u and v in practice. However it must be noted that for sufficiently large n, any one $\zeta_{u,v}$ is estimable.

It is also to be noted that Theorem 2, and therefore Corollary 1, exclude the case when the underlying multinomial distribution is uniform. This exclusion makes the asymptotic normality somewhat incomplete. However this should not be taken as if $Z_{u,v}$ is less of an estimator in that excluded case. On the contrary, $Z_{u,v}$ in this case is sometimes called a super efficient estimator with a variance degenerating faster than $n^{-1/2}$. The asymptotic distribution of a properly normalized $Z_{u,v}$ exists and can be derived, but it would have little or no practical value and therefore is omitted from this chapter.

Definition 2. A multi-dimensional parameterization of an underlying distribution, $\{\theta\} \in \Theta$, is said to be sufficient iff $\{\theta\}$ uniquely determines the underlying distribution.

Definition 3. A multi-dimensional parameterization of an underlying distribution, $\{\theta\} = \{\theta_\beta; \beta \in B\} \in \Theta$ for some index set B, is said to be minimally sufficient iff 1) $\{\theta\}$ is sufficient; and 2) there does not exist a proper subset of B, $B' \subset B$, such that $\{\theta\}' = \{\theta_\beta; \beta \in B'\}$ is sufficient.

Definition 4. Two multi-dimensional parameterizations of an underlying distribution, $\{\theta\} \in \Theta$ and $\{\omega\} \in \Omega$, are said to be equivalent, denoted by $\{\theta\} \rightleftharpoons \{\omega\}$, iff an one-to-one mapping from Θ to Ω exists.

For the family of infinite dimensional multinomial distributions $\{p_s\}$, $\{p_s; s \geq 1\}$ is sufficient but not minimally sufficient since $\{p_s; s \geq 2\}$ is also sufficient. In fact, $\{p_s; s \geq$

$1, s \neq s_0\}$ for any $s_0 \geq 1$ is minimally sufficient; and $\{p_s; s \geq 1, s \neq s_1, s \neq s_2, s_1 \neq s_2\}$ for any $s_1 \geq 1$ and $s_2 \geq 1$ is not sufficient. $\{p_s^\alpha; s \geq 1, s \neq s_0\}$ for any fixed $\alpha > 0$ is also minimally sufficient.

By Theorem 1, under P', $\{\zeta_u; u \geq 1\} \rightleftharpoons \{p_s; s \geq 1\}$. Since $\{\zeta_u; u \geq 1\} \subset \{\zeta_{u,v}; u \geq 1, v \geq 0\}$, $\{\zeta_{u,v}; u \geq 1, v \geq 0\} \rightleftharpoons \{p_s; s \geq 1\}$. Similarly since $\{\mathcal{N}_\alpha; \alpha \geq 0\} \rightleftharpoons \{(\mathcal{N}_\alpha)^{1-\alpha}; \alpha \geq 0\}$ and $\{\zeta_u; u \geq 1\} \subset \{(\mathcal{N}_\alpha)^{1-\alpha}; \alpha \geq 0\}$, $\{\mathcal{N}_\alpha; \alpha \geq 0\} \rightleftharpoons \{p_s; s \geq 1\}$. This is to say that both the generalized Simpson's indices and the family of the Rényi-Hill indices are sufficient.

On the other hand, $\{\zeta_u; u \geq 1\}$ is not minimally sufficient, which implies that $\{\mathcal{N}_\alpha; \alpha \geq 0\}$ is not minimally sufficient. The fact that $\{\zeta_u; u \geq 1\}$ is not minimally sufficient can be seen by the fact that any subsequence of $\{\zeta_u\}$ uniquely determines the underlying distribution. The proof of that fact is identical to that of Theorem 1. Furthermore and more interestingly, a minimally sufficient subsequence of $\{\zeta_u\}$ does not exist, since a subsequence of any subsequence will uniquely determine the underlying distribution.

CHAPTER 2: ASYMPTOTIC PROPERTIES OF TWO DIMENSIONAL SAMPLE COVERAGE ESTIMATORS

2.1 Introduction

Consider a multinomial distribution with countably infinite number of categories indexed by $K = \{k; k = 1, 2, \cdots\}$ and category probabilities denoted by $\{p_k\}$, satisfying $0 < p_k < 1$ for all k and $\sum p_k = 1$, where the sum without index is over all k as in all subsequent text of this chapter unless otherwise stated. (In fact, in the subsequent text of this chapter, we should observe the convention that $\sum_{K_i} = \sum_{k \in K_i}$, $\prod_{K_i} = \prod_{k \in K_i}$, $\lim = \lim_{n \to \infty}$ and that $\int = \int_{-\infty}^{+\infty}$, unless otherwise indicated. We also use "$\sim$" to indicate equality in the limit.) Denote the category counts in an *iid* sample of size n from that population by $(x_1, \cdots)$. Note that for a given sample, there are at most n non-zero x_k's. Suppose the target of estimation is the "total probability of the categories not represented in the sample", or equivalently

$$\pi_0 = \sum p_k I[x_k = 0] \tag{1}$$

where $I[\cdot]$ is the indicator function. It may be interesting to note that π_0 is not a fixed constant nor is it an observable random variable. This target is interesting because it represents the probability that the $(n+1)$th observation is from a previously unobserved category.

An estimate described by Good (1953), but largely credited to Turing and hence known as Turing's formula, is given by

$$T = \frac{N_1}{n} \tag{2}$$

where N_1 is the number of categories represented exactly once in the sample, *i.e.*, $N_1 = \sum I[x_k = 1]$. This simple formula has been used widely across many fields of study, frequently in the form of $C' = 1 - T$ estimating $C = 1 - \pi_0$ which is often referred to as the "coverage" problem.

Many authors have discussed issues related to this problem in various settings. However its asymptotic normality was not known for a long time until Esty (1983) who gave a set of conditions for a $\sqrt{n}$-normalized convergence theorem for the case when $\{p_k\}$ is changing with respect to sample size n. After 25 years since Esty, in 2008, Zhang and Huang derived the asymptotic normality for the case when $\{p_k\}$ is fixed with respect to sample size n.

2.2 Motivation

Since Turing's formula is an asymptotic unbiased estimator of π_0. And π_0 characterize the tail probability when the sample size increases. It is natural to link the Turing's formula with the problems about tail probability. In 1975, Hill proposed a simple general approach to make inference about the tail behavior of a distribution. It is not required to assume any global form for the distribution function, but merely the parametric form of behavior in the tail. However, Hill's estimator is correct only for very large values in the sample. But how large the values should be in order to make the estimator be valid? So far, there is no clear answer in the literature. There are two possible ways can solve this problem. Either we can try to find a way to determine this boundary value or we can avoid to determine this boundary value explicitly. Since Turing's formula exactly characterize the tail behavior of a distribution when the sample size increases, it is natural to link the Turing's formula with Hill's approach. The major advantage of this new approach is that we do not need to explicitly determine how large the values should be in order to make Hill's estimator be valid. However there are two parameters need to be estimated in Hill's approach. The current one dimensional asymptotic property of Turing's formula which derived by Zhang and Huang (2008) is not enough to acquire the estimation of Hill's approach. Therefore high dimensional asymptotic results are expected. Motivated by this consideration, we derived a two-dimensional asymptotic normality for Turing's formula under certain conditions.

We split the categories from the original population into two sub-categories corresponding to two sub-populations with infinite categories in each of them. Let the first and second sub-populations with countable infinite categories indexed by K_1 and K_2 respectively where $K_1 \subset K$, $K_2 \subset K$ and $K_2 = K \backslash K_1$. Suppose the targets of estimation are the "total prob-

abilities of the categories of sub-populations not represented in the sample", or equivalently

$$\pi_i = \sum_{K_i} p_k I[x_k = 0], \qquad i = 1, 2. \tag{3}$$

According to Turing's formula, we define

$$T_1 = \frac{N_1}{n}, \qquad T_2 = \frac{M_1}{n} \tag{4}$$

where N_1 and M_1 are the number of categories in the first and second sub-populations represented exactly once in the sample, *i.e.*, $N_1 = \sum_{K_1} I[x_k = 1]$ and $M_1 = \sum_{K_2} I[x_k = 1]$.

Motivated by the Zhang and Huang (2008), we derived two dimensional asymptotic normality for $\mathbf{Z} = (Z_1, Z_2)'$ where $Z_i = \pi_i - T_i,\ i = 1, 2$.

2.3 Asymptotic Results

Let $K = \{k; k = 1, \cdots\}$ be the index set of all the positive integers. K_1 and K_2 are two subsets of K with infinite elements in each of them and $K_1 = K \backslash K_2$. Let

$$f_k(x) = \begin{cases} p_k & x = 0, \\ -1/n & x = 1, \\ 0 & x \geq 2. \end{cases}$$

$Z_i = \sum_{K_i} f_k(X_k) = \pi_i - T_i,\ i = 1, 2$. We are interested in the asymptotic behavior of $\mathbf{Z}g(n)$, where $g(n)$ is a function of n satisfying $\lim_{n\to\infty} g(n) = \infty$ and

$$g(n) = O(n^{1-2\delta}) \tag{5}$$

for some $\delta \in (0, 1/4)$. $\mathbf{Z} = (Z_1, Z_2)'$.

In order to acquire the asymptotic normality of this two-dimensional random vector, we need to show that any linear combination of elements of this vector asymptotically follows one dimensional normal distribution as $n \to \infty$.

For any real constants, a and b, satisfying $a^2 + b^2 \neq 0$, consider

$$\begin{aligned} Z &= aZ_1 + bZ_2 \\ &= a\sum_{k\in K_1} f_k(x_k) + b\sum_{k\in K_2} f_k(x_k) \\ Zg(n) &= ag(n)\sum_{k\in K_1} f_k(x_k) + bg(n)\sum_{k\in K_2} f_k(x_k) \end{aligned}$$

We are interested in the asymptotic behavior of $Zg(n)$ in terms of the limit of its characteristic function, $E[\exp(isZg(n))]$.

Lemma 2. Let $\{X_k\}$ be the counts of observations in category k, $k = 1, 2, \cdots$, in an *iid* sample under the multinomial model with probability distribution $\{p_k\}$. Then

$$P(X_k = x_k; k = 1, 2, \cdots) = P(Y_k = x_k; k = 1, 2, \cdots | \sum Y_k = n)$$

where Y_k are independent possion random variables with mean np_k.

Lemma 3. Let (U, V) be a two-dimensional random vector with U integer valued. Then

$$E(\exp(ivV|U = u)) = (2\pi P(U = n))^{-1} \int_{-\pi}^{\pi} E(\exp(iu(U - n) + ivV))du.$$

The Lemma 2 is a well-known fact and lemma 3 is due to Bartlett (1938). Based on these two lemmas,

$$E(\exp(isZg(n))) = (2\pi P(\sum_K Y_k = n))^{-1} \int_{-\pi}^{\pi} E[\exp(iu\sum_K (Y_k - np_k) + isZg(n))]du.$$

We want to evaluate $\lim_{n\to\infty} E(\exp(isZg(n)))$. Toward this end, we first note that, by Stirling's formula, $(2\pi n)^{1/2} P(\sum_K Y_k = n) \to 1$. Therefore we need only to evaluate the limit of

$$H_n(s) = \frac{\sqrt{n}}{\sqrt{2\pi}} \int_{-\pi}^{\pi} E[\exp(iu\sum_K (Y_k - np_k) + isZg(n))]du, \tag{6}$$

or letting $t = un^{1/2}$,

$$H_n(s) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} I[|t| < \pi\sqrt{n}] E[\exp(i(n)^{-1/2} t \sum_K (Y_k - np_k) + isZg(n))] dt. \quad (7)$$

Writing

$$\begin{aligned} h_n &= I[|t| < \pi\sqrt{n}] E[\exp(i(n)^{-1/2} t \textstyle\sum_K (Y_k - np_k) + isZg(n))] \\ &= I[|t| < \pi\sqrt{n}] E[\exp(i(n)^{-1/2} t \textstyle\sum_{K_1} (Y_k - np_k) + isag(n) \sum_{K_1} f_k(Y_k) \\ &\quad + i(n)^{-1/2} t \textstyle\sum_{K_2} (Y_k - np_k) + isbg(n) \sum_{K_2} f_k(Y_k))] \end{aligned} \quad (8)$$

According to lemma 2, $\{Y_k\}$ are independent Poisson random variables with mean np_k. Therefore,

$$\begin{aligned} h_n &= I[|t| < \pi\sqrt{n}] E[\exp(i(n)^{-1/2} t \textstyle\sum_{K_1} (Y_k - np_k) + isag(n) \sum_{K_1} f_k(Y_k))] \\ &\quad \times E[\exp(i(n)^{-1/2} t \textstyle\sum_{K_2} (Y_k - np_k) + isbg(n) \sum_{K_2} f_k(Y_k))]. \end{aligned} \quad (9)$$

Let,

$$\begin{aligned} A_1 &= E[\exp(i(n)^{-1/2} t \textstyle\sum_{K_1} (Y_k - np_k) + isag(n) \sum_{K_1} f_k(Y_k))] \\ A_2 &= E[\exp(i(n)^{-1/2} t \textstyle\sum_{K_2} (Y_k - np_k) + isbg(n) \sum_{K_2} f_k(Y_k))] \end{aligned} \quad (10)$$

Our first task is to allow the limit operator to exchange with the integral operator. By definition of A_1 and A_2, $h_n = I[|t| < \pi\sqrt{n}] A_1 \times A_2$. Let's define two index sets K_{11} and K_{12} where K_{11} only contains one element from K_1, let's call it r, and $K_{12} = K_1 \backslash K_{11}$. Writing

$$\begin{aligned} A_{11} &= I[|t| < \pi\sqrt{n}] E[\exp(i(n)^{-1/2} t (Y_r - np_r) + isaf_r(Y_r) g(n)] \\ A_{12} &= I[|t| < \pi\sqrt{n}] E[\exp(i(n)^{-1/2} t \textstyle\sum_{K_{12}} (Y_k - np_k) + isa \sum_{K_{12}} f_k(Y_k) g(n))], \end{aligned} \quad (11)$$

$$H_n(s) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} h_n dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} I[|t| < \pi\sqrt{n}] A_{11} A_{12} A_2 dt. \tag{12}$$

Since $|A_2| \leq 1$ and $|A_{12}| \leq 1$, $|h_n| \leq |A_{11}|$. On the other hand,

$$E[\exp(iu(Y_r - np_r) + isaf_r(Y_r)g(n))]$$

$$= \exp(iu(-np_r) + isap_r g(n)) \exp(-np_r) + \exp(iu(1 - np_r) - isan^{-1}g(n)) np_r \exp(-np_r)$$

$$+ \sum_{j=2}^{\infty} \exp(iu(j - np_r)) P(Y_r = j)$$

$$= \sum_{j=0}^{\infty} \exp(iu(j - np_r)) P(Y_r = j)$$

$$- \exp(-iunp_r) \exp(-np_r) - \exp(iu(1 - np_r)) np_r \exp(-np_r)$$

$$+ \exp(iu(-np_r) + isap_r g(n)) \exp(-np_r) + \exp(iu(1 - np_r) - n^{-1} isag(n)) np_r \exp(-np_r)$$

$$= [\exp(-iunp_r) \exp(i \sin(u) np_r) \exp(np_r(\cos(u) - 1))]$$

$$- \exp(-iunp_r) \exp(-np_r) - \exp(iu(1 - np_r)) np_r \exp(-np_r)$$

$$+ \exp(iu(-np_r) + isap_r g(n)) \exp(-np_r) + \exp(iu(1 - np_r) - n^{-1} isag(n)) np_r \exp(-np_r).$$

Therefore (recall $t = u\sqrt{n}$),

$$|A_{11}| \leq I[|t| < \pi\sqrt{n}] \left[\exp(np_r(\cos(tn^{-\frac{1}{2}}) - 1)) + 2[\exp(-np_r) + np_r \exp(-np_r)] \right] \qquad (= \overline{A}_{11}).$$

It is clear that, for any t, by Taylor's formula for $\cos(x)$,

$$\lim_{n\to\infty} \overline{A}_{11} = \lim_{n\to\infty} I[|t| < \pi\sqrt{n}] \exp(np_r(\cos(tn^{-1/2}) - 1)) = \exp(-p_r t^2/2) \qquad (= \overline{A}_1).$$

$$\begin{aligned}
\int_{-\infty}^{+\infty} |\overline{A}_{11}| dt &= \int_{-\infty}^{+\infty} I[|t| < \pi\sqrt{n}] \left[\exp(np_r(\cos(tn^{-\frac{1}{2}}) - 1))\right] dt \\
&\quad +2\int_{-\infty}^{+\infty} I[|t| < \pi\sqrt{n}] \exp(-np_r) dt + 2\int_{-\infty}^{+\infty} I[|t| < \pi\sqrt{n}] np_r \exp(-np_r) dt \\
&= \int_{-\infty}^{+\infty} I[|t| < \pi\sqrt{n}] \left[\exp(np_r(\cos(tn^{-\frac{1}{2}}) - 1))\right] dt \\
&\quad +2 \times 2\pi\sqrt{n} \exp(-np_r) + 2 \times 2\pi\sqrt{n} np_r \exp(-np_r).
\end{aligned}$$

Since the last two terms vanish to zero as $n \to \infty$, we have, letting δ be a constant in $(0, 1/2)$,

$$\begin{aligned}
\lim_{n\to\infty} \int_{-\infty}^{+\infty} |\overline{A}_{11}| dt &= \lim_{n\to\infty} \int_{-\infty}^{+\infty} I[|t| < \pi\sqrt{n}] \left[\exp(np_r(\cos(tn^{-\frac{1}{2}}) - 1))\right] dt \\
&= \lim_{n\to\infty} \int_{-\pi}^{+\pi} \sqrt{n} \left[\exp(np_r(\cos(u) - 1))\right] du \\
&= \lim_{n\to\infty} \int_{|u| < \frac{1}{n^{(1-\delta)/2}}} \sqrt{n} \left[\exp(np_r(\cos(u) - 1))\right] du \\
&\quad + \lim_{n\to\infty} \int_{\frac{1}{n^{(1-\delta)/2}} \le |u| < \pi} \sqrt{n} \left[\exp(np_r(\cos(u) - 1))\right] du \\
&(= \lim_{n\to\infty} \eta_1 + \lim_{n\to\infty} \eta_2).
\end{aligned}$$

The second term of the last expression above is zero. To see this, we note that for any

u satisfying $\frac{1}{n^{(1-\delta)/2}} \leq |u| < \pi$, $\cos(u) - 1 \leq \cos(1/n^{(1-\delta)/2}) - 1$, and hence

$$\begin{aligned}
\lim_{n\to\infty} \eta_2 &\leq \lim_{n\to\infty} \int_{\frac{1}{n^{(1-\delta)/2}} \leq |u| < \pi} \sqrt{n} \left[\exp(np_r(\cos(1/n^{(1-\delta)/2}) - 1))\right] du \\
&= \lim_{n\to\infty} 2\pi\sqrt{n} \left[\exp(np_r(\cos(1/n^{(1-\delta)/2}) - 1))\right] \\
&= \lim_{n\to\infty} 2\pi\sqrt{n} \left[\exp\left(-np_r\left(1 - \cos(1/n^{(1-\delta)/2})\right)\right)\right] \\
&= \lim_{n\to\infty} 2\pi\sqrt{n} \left[\exp\left(-np_r\left(\frac{\sin^2(1/n^{(1-\delta)/2})}{1+\cos(1/n^{(1-\delta)/2})}\right)\right)\right] \\
&= \lim_{n\to\infty} 2\pi\sqrt{n} \exp\left(-np_r O(\tfrac{1}{n^{1-\delta}})\right) \\
&= \lim_{n\to\infty} 2\pi\sqrt{n} \exp\left(-p_r O(n^{\delta})\right) = 0.
\end{aligned}$$

For u satisfying $|u| < \frac{1}{n^{(1-\delta)/2}}$, consider the Taylor expansion of

$$\begin{aligned}
\cos(u) - 1 &= -\tfrac{u^2}{2!} + \tfrac{u^4}{4!} - \tfrac{u^6}{6!} + \cdots + \tfrac{(-1)^m u^{2m}}{(2m)!} + \cdots \\
&\leq -\tfrac{u^2}{2} + (u^4 + u^8 + \cdots + u^{4m} + \cdots) \\
&= -\tfrac{u^2}{2} + \tfrac{u^4}{1-u^4}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\lim_{n\to\infty}\eta_2 &\leq \lim_{n\to\infty}\int_{|u|<\frac{1}{n^{(1-\delta)/2}}}\sqrt{n}\exp\left(np_r\left(-\frac{u^2}{2}+\frac{\frac{1}{n^{2-2\delta}}}{1-\frac{1}{n^{2-2\delta}}}\right)\right)du \\
&= \lim_{n\to\infty}\int_{|u|<\frac{1}{n^{(1-\delta)/2}}}\sqrt{n}\exp\left(-\frac{np_ru^2}{2}+np_r\frac{\frac{1}{n^{2-2\delta}}}{1-\frac{1}{n^{2-2\delta}}}\right)du \\
&= \lim_{n\to\infty}\left[\left(\int_{|u|<\frac{1}{n^{(1-\delta)/2}}}\sqrt{n}\exp\left(-\frac{np_ru^2}{2}\right)du\right)\exp\left(O(\tfrac{1}{n^{1-2\delta}})\right)\right] \\
&\quad (\text{ letting } t=u\sqrt{n}) \\
&= \lim_{n\to\infty}\left[\left(\int_{|t|<n^{\delta/2}}\exp\left(-\frac{p_rt^2}{2}\right)dt\right)\exp\left(O(\tfrac{1}{n^{1-2\delta}})\right)\right] \\
&= \int_{-\infty}^{+\infty}\exp\left(-p_rt^2/2\right)dt.
\end{aligned}$$

Since $\cos(u)\geq -\frac{u^2}{2}$ for all u satisfying $|u|<\frac{1}{n^{(1-\delta)/2}}$, it is easy to establish $\lim_{n\to\infty}\eta_2\geq \int_{-\infty}^{+\infty}\exp\left(-p_rt^2/2\right)dt$, and hence $\lim_{n\to\infty}\eta_2=\int_{-\infty}^{+\infty}\exp\left(-p_rt^2/2\right)dt$.

Now that we have established

$$\lim_{n\to\infty}\int_{-\infty}^{+\infty}|\overline{A}_{11}|dt=\int_{-\infty}^{+\infty}\lim_{n\to\infty}|\overline{A}_{11}|dt,$$

by the Dominated Convergence Theorem. We have the following lemma.

Lemma 4. Let h_n and H_n be as defined in (7) and (8) respectively. Then

$$\lim_{n\to\infty}H_n=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\lim_{n\to\infty}h_ndt.$$

We now turn to evaluate $\lim h_n$. Since $h_n=I[|t|<\pi\sqrt{n}]A_1\times A_2$, we will firstly evaluate A_1 and A_2 seperately. For each $k_1\in K_1$ and $k_2\in K_2$, it can be verified that,

letting

$$
\begin{aligned}
B_{k_1} &= \exp(-itp_{k_1}n^{1/2})[\exp(np_{k_1}(\exp(itn^{-1/2})-1))] \\
C_{k_1} &= \exp(-itp_{k_1}n^{1/2})[\exp(isap_{k_1}g(n))-1]\exp(-np_{k_1}) \\
D_{k_1} &= \exp(-itp_{k_1}n^{1/2})\exp(itn^{-1/2})[\exp(-isan^{-1}g(n))-1]np_{k_1}\exp(-np_{k_1}) \\
B'_{k_2} &= \exp(-itp_{k_2}n^{1/2})[\exp(np_{k_2}(\exp(itn^{-1/2})-1))] \\
C'_{k_2} &= \exp(-itp_{k_2}n^{1/2})[\exp(isbp_{k_2}g(n))-1]\exp(-np_{k_2}) \\
D'_{k_2} &= \exp(-itp_{k_2}n^{1/2})\exp(itn^{-1/2})[\exp(-isbn^{-1}g(n))-1]np_{k_2}\exp(-np_{k_2}) \\
E_{k_1} &= C_{k_1}+D_{k_1} \\
E'_{k_2} &= C'_{k_2}+D'_{k_2},
\end{aligned}
\tag{13}
$$

then, $A_1 = \prod_{K_1}(B_k+E_k)$ and $A_2 = \prod_{K_2}(B'_k+E'_k)$. And $h_n \sim \prod_{K_1}(B_k+E_k)\prod_{K_2}(B'_k+E'_k)$. We are interested in evaluating $\lim\prod_{K_1}(B_k+E_k)$ and $\lim\prod_{K_2}(B'_k+E'_k)$.

The facts of the following two lemmas are given by Esty (1983).

Lemma 5. Let $\{\beta_k\}$ and $\{\epsilon_k\}$ be two sequences of complex numbers, and M_n be a sequence of subsets of K, indexed by n. If

1. $\prod_{M_n}\beta_k \sim \beta$,
2. $(\sum_{M_n}\epsilon_k) \sim \epsilon$,
3. $\beta_k \sim 1$ uniformly,
4. $\epsilon_k \sim 0$ uniformly,
5. there exists a constant, δ_1 such that, $\sum_{M_n}|\beta_k-1| \leq \delta_1$,

6. there exists a constant, δ_2 such that, $\sum_{M_n} |\epsilon_k| \leq \delta_2$;

then

$$\prod_{M_n} (\beta_k + \epsilon_k) \sim \beta e^{\epsilon}$$

where β and ϵ may also depend on n.

Lemma 6. For all $k \in K$, $B_k = \exp[(-t^2/2)p_k + O(t^3 p_k n^{-1/2})]$.

The next lemma includes three useful facts.

Lemma 7. 1. For any complex number x satisfying $|x| < 1$, $|\ln(1+x)| \leq \frac{|x|}{1-|x|}$.

2. For any real number $x \in [0,1)$, $1 - x \geq \exp(-\frac{x}{1-x})$.

3. For any real number $x \in (0, 1/2)$, $\frac{1}{1-x} < 1 + 2x$.

Let us consider partitions of the index sets $K_1 = I_1 \bigcup II_1$ and $K_2 = I_2 \bigcup II_2$

$$I_1 = \{k : k \in K_1, p_k g(n) \leq n^{-\delta}\} \quad \text{and} \quad II_1 = \{k : k \in K_1, p_k g(n) > n^{-\delta}\}$$

$$I_2 = \{k : k \in K_2, p_k g(n) \leq n^{-\delta}\} \quad \text{and} \quad II_2 = \{k : k \in K_2, p_k g(n) > n^{-\delta}\}$$

where δ is as in (5).

Lemma 8. (a) $\sum_{II_1} |E_k| \to 0$ and (b) $\prod_{II_1}(B_k + E_k) / \prod_{II_1} B_k \to 1$.

Proof. (a) $\sum_{II_1} |E_k| \leq 2 \sum_{II_1} (e^{-np_k} + np_k e^{-np_k})$. Since the derivative of $(e^{-np_k} + np_k e^{-np_k})$ for any $k \in II_1$ is negative with respect to p_k. It attains its maximum at $p_k = 1/(g(n)n^{\delta})$ with value $e^{-n/(g(n)n^{\delta})}(1 + n/(g(n)n^{\delta}))$. The total number of indices in II_1 is less than or equals to $g(n)n^{\delta}$. Therefore,

$$\sum_{II_1} |E_k| \leq 2 \left(g(n)n^{\delta}\right) \left(e^{-n/(g(n)n^{\delta})}(1 + n/(g(n)n^{\delta}))\right) = 2e^{-O(n^{\delta})} O(n) \to 0$$

(b) By lemma 6, $|B_k|$ is bounded away from zero, and by the fact that $\lim |E_k| = 0$ and by applying the first part of lemma 7 with $x = E_k/B_k$, we can get

$$\begin{aligned} \left|\ln[\textstyle\prod_{II_1}(B_k + E_k)/\prod_{II_1} B_k]\right| &= \left|\textstyle\sum_{II_1} \ln(1 + E_k/B_k)\right| \leq \textstyle\sum_{II_1} |\ln(1 + E_k/B_k)| \\ &\leq \textstyle\sum_{II_1} \left(\frac{|E_k|}{|B_k|-|E_k|}\right) = O(\textstyle\sum_{II_1} |E_k|) \to 0. \end{aligned} \tag{14}$$

The following conditions are the sufficient conditions to get many subsequent results.

Condition 2.3.1. As $n \to \infty$,

1. $\sum_{K_1}(g^2(n)/n)p_k e^{-np_k} \to c_1 \geq 0$,
2. $\sum_{K_1} g^2(n)p_k^2 e^{-np_k} \to c_2 \geq 0$,
3. $c_1 + c_2 > 0$,
4. $\sum_{K_2}(g^2(n)/n)p_k e^{-np_k} \to d_1 \geq 0$,
5. $\sum_{K_2} g^2(n)p_k^2 e^{-np_k} \to d_2 \geq 0$, and
6. $d_1 + d_2 > 0$.

Lemma 9. Under Condition 2.3.1, all the conditions of lemma 5 are satisfied with $M_n = I_1$, $\beta_k = B_k$, $\beta = B$, $\epsilon_k = E_k$ and $\epsilon = E$.

Proof. We need to check all six conditions in Lemma 5.

For 3), it is true because from lemma 6, $B_k = \exp[(-t^2/2)p_k + O(t^3 p_k n^{-1/2})]$, and p_k, $p_k/\sqrt{n}$ are uniformly bounded by $\frac{1}{g(n)n^\delta}$ and $\frac{1}{g(n)\sqrt{n}n^\delta}$ respectively. As $n \to 0$, $B_k \sim 1$ uniformly in M_n.

For 1), since $\sum_{I_1} p_k \to 0$,

$$\prod_{I_1} B_k = \exp(-(t^2/2)\sum_{I_1} p_k)\exp(O((t^3/n^{-1/2})\sum_{I_1} p_k) \to 1.$$

For 2), 4) and 6),

$$
\begin{aligned}
E_{k_1} &= e^{-np_{k_1}}e^{-itp_{k_1}\sqrt{n}}\{isag(n)p_{k_1} - \tfrac{s^2a^2g^2(n)p_{k_1}^2}{2} + O(s^3a^3g^3(n)p_{k_1}^3) \\
&\quad + np_{k_1}[1 + \tfrac{it}{\sqrt{n}} - \tfrac{t^2}{2n} + O(\tfrac{t^3}{n^{3/2}})][-\tfrac{isag(n)}{n} - \tfrac{s^2a^2g^2(n)}{2n^2} + O(\tfrac{s^3a^3g^3(n)}{n^3})]\} \\
&= e^{-np_{k_1}}e^{-itp_{k_1}\sqrt{n}}\{isag(n)p_{k_1} - \tfrac{s^2a^2g^2(n)p_{k_1}^2}{2} + O(s^3a^3g^3(n)p_{k_1}^3) \\
&\quad + [np_{k_1} + itp_{k_1}\sqrt{n} - \tfrac{t^2p_{k_1}}{2} + np_{k_1}O(\tfrac{t^3}{n^{3/2}})][-\tfrac{isag(n)}{n} - \tfrac{s^2a^2g^2(n)}{2n^2} + O(\tfrac{s^3a^3g^3(n)}{n^3})]\} \\
&= e^{-np_{k_1}}e^{-itp_{k_1}\sqrt{n}}\{isag(n)p_{k_1} - \tfrac{s^2a^2g^2(n)p_{k_1}^2}{2} + O(s^3a^3g^3(n)p_{k_1}^3) \\
&\quad - isag(n)p_{k_1} - \tfrac{s^2a^2}{2}(\tfrac{g^2(n)p_{k_1}}{n}) + np_{k_1}O(\tfrac{s^3a^3g^3(n)}{n^3}) \\
&\quad + sta\tfrac{g(n)}{\sqrt{n}}p_{k_1} - \tfrac{is^2ta^2}{2n^{3/2}}g^2(n)p_{k_1} + itp_{k_1}\sqrt{n}O(\tfrac{s^3a^3g^3(n)}{n^3}) \\
&\quad + \tfrac{ist^2a}{2}\tfrac{g(n)}{n}p_{k_1} + \tfrac{s^2t^2a^2}{4}\tfrac{g^2(n)}{n^2}p_{k_1} - \tfrac{t^2p_{k_1}}{2}O(\tfrac{s^3a^3g^3(n)}{n^3}) \\
&\quad - isag(n)p_{k_1}O(\tfrac{t^3}{n^{3/2}}) - \tfrac{s^2a^2}{2}\tfrac{g^2(n)}{n}p_{k_1}O(\tfrac{t^3}{n^{3/2}}) + np_{k_1}O(\tfrac{t^3}{n^{3/2}})O(\tfrac{s^3a^3g^3(n)}{n^3})\} \\
&= e^{-np_{k_1}}e^{-itp_{k_1}\sqrt{n}}\{-\tfrac{s^2a^2g^2(n)p_{k_1}^2}{2} - \tfrac{s^2a^2}{2}(\tfrac{g^2(n)p_{k_1}}{n}) \\
&\quad + sta\tfrac{g(n)}{\sqrt{n}}p_{k_1} + \tfrac{s^2t^2a^2}{4}\tfrac{g^2(n)}{n^2}p_{k_1} - \tfrac{is^2ta^2}{2n^{3/2}}g^2(n)p_{k_1} + \tfrac{ist^2a}{2}\tfrac{g(n)}{n}p_{k_1} \\
&\quad + O(s^3a^3g^3(n)p_{k_1}^3) + O(\tfrac{s^3a^3g^3(n)}{n^2}p_{k_1}) + iO(ts^3a^3\tfrac{g^3(n)}{n^{5/2}}p_{k_1}) - O(\tfrac{s^3t^2a^3}{2}\tfrac{g^3(n)}{n^3}p_{k_1}) \\
&\quad - iO(st^3a\tfrac{g(n)}{n^{3/2}}p_{k_1}) - O(\tfrac{s^2t^3a^2}{2}\tfrac{g^2(n)}{n^{5/2}}p_{k_1} + O(\tfrac{s^3a^3t^3}{2}\tfrac{g^3(n)}{n^{7/2}}p_{k_1})\}.
\end{aligned}
\tag{15}
$$

For all $k \in I_1$, $e^{-itp_k\sqrt{n}} \to 1$ uniformly since $p_k\sqrt{n} \le \frac{\sqrt{n}}{g(n)n^\delta}$. And it is easy to check that every additive term in E_k converges to zero uniformly for all $k \in I_1$. Therefore (4) is

checked.

It is easy to check that for every term within the curly brackets in (15), denoted by $\tau(s,t,n,p_k)$, except the first two terms,

$$\left|\sum_{I_1} e^{-np_k}\tau(s,t,n,p_k)\right| \le \sum_{I_1} e^{-np_k}|\tau(s,t,n,p_k)| \to 0$$

uniformly by Condition 2.3.1.

The uniform convergence of $\sum_{I_1} e^{-np_k} g^2(n) p_k^2$ and $\sum_{I_1} e^{-np_k} \frac{g^2(n)}{n} p_k$ are directly guaranteed by Condition 2.3.1. Therefore (2) is checked. The uniformity of convergence for $\sum_{I_1} E_k$ and hence for $\sum_{I_1} |E_k|$ guarantees (6).

For 5), since $B_k = \exp[(-t^2/2)p_k + O(t^3 p_k n^{-1/2})]$ and $(-t^2/2)p_k + O(t^3 p_k n^{-1/2}) \to 0$ uniformly, we have

$$|B_k - 1| \le \frac{|(-t^2/2)p_k + O(t^3 p_k n^{-1/2})|}{1 - |(-t^2/2)p_k + O(t^3 p_k n^{-1/2})|} = O((-t^2/2)p_k + t^3 p_k n^{-1/2})$$

and hence

$$\sum_{I_1} |B_k - 1| \le O(\frac{t^2}{2}\sum_{I_1} p_k + \frac{|t^3|}{\sqrt{n}}\sum_{I_1} p_k) < O(t^2 + |t^3|).$$

Corollary 2. Under Condition 2.3.1, all the conditions of lemma 5 are satisfied with $M_n = I_2$, $\beta_k = B_k^{'}$, $\beta = B^{'}$, $\epsilon_k = E_k^{'}$ and $\epsilon = E^{'}$.

The proof of corollary 2 is similar to the proof of lemma 9 except changing the I_1, B_k, B, E_k and E to I_2, $B_k^{'}$, $B^{'}$, $E_k^{'}$ and $E^{'}$ respectively.

Corollary 3. Under Condition 2.3.1, $\prod_{I_1}(B_k + E_k) \sim \prod_{I_1} B_k \exp(\sum_{I_1} E_k)$ and $\prod_{I_2}(B_k^{'} + E_k^{'}) \sim \prod_{I_2} B_k^{'} \exp(\sum_{I_2} E_k^{'})$.

Lemma 10. Under Condition 2.3.1, $\prod_{K_1}(B_k + E_k) \to Be^E$, where $B = \lim \prod_{K_1} B_k$, $E = \lim \sum_{K_1} E_k$ and $\prod_{K_2}(B_k^{'} + E_k^{'}) \to B^{'} e^{E^{'}}$, where $B^{'} = \lim \prod_{K_2} B_k^{'}$, $E^{'} = \lim \sum_{K_2} E_k^{'}$.

Proof.

$$\begin{aligned}
\prod_{K_1}(B_k+E_k) &= \prod_{I_1}(B_k+E_k)\prod_{II_1}(B_k+E_k)\\
&\sim \prod_{I_1}(B_k+E_k)\prod_{II_1}B_k\\
&\sim \prod_{I_1}B_k\exp(\textstyle\sum_{I_1}E_k)\prod_{II_1}(B_k)\\
&\sim \prod_{K_1}B_k\exp(\textstyle\sum_{K_1}E_k).\\
\prod_{K_2}(B_k^{'}+E_k^{'}) &= \prod_{I_2}(B_k^{'}+E_k^{'})\prod_{II_2}(B_k^{'}+E_k^{'})\\
&\sim \prod_{I_2}(B_k^{'}+E_k^{'})\prod_{II_2}B_k^{'}\\
&\sim \prod_{I_2}B_k^{'}\exp(\textstyle\sum_{I_2}E_k^{'})\prod_{II_2}(B_k^{'})\\
&\sim \prod_{K_2}B_k^{'}\exp(\textstyle\sum_{K_2}E_k^{'}).
\end{aligned} \tag{16}$$

Theorem 3. Let $g(n)$ be as in (5). Under condition 2.3.1,

$$g(n)Z \xrightarrow{d} N\left(0,(a^2(c_1+c_2)+b^2(d_1+d_2))\right).$$

Proof. Since $\lim \prod B_k = e^{-\frac{t^2}{2}}$ and

$$\begin{aligned}
\lim \textstyle\sum_{K_1} E_k &= -\tfrac{s^2a^2}{2}(\lim \textstyle\sum_{K_1} \tfrac{g^2(n)}{n} p_k e^{-np_k} + \lim \textstyle\sum_{K_1} g^2(n) p_k e^{-np_k}) \\
\lim \textstyle\sum_{K_2} E_k' &= -\tfrac{s^2b^2}{2}(\lim \textstyle\sum_{K_2} \tfrac{g^2(n)}{n} p_k e^{-np_k} + \lim \textstyle\sum_{K_2} g^2(n) p_k e^{-np_k}) \\
\lim H_n &= (\tfrac{1}{\sqrt{2\pi}} \textstyle\int e^{-\frac{t^2}{2}} dt) \exp(-\tfrac{s^2a^2}{2}(\lim \textstyle\sum_{K_1} \tfrac{g^2(n)}{n} p_k e^{-np_k} + \lim \textstyle\sum_{K_1} g^2(n) p_k e^{-np_k}) \\
&\quad -\tfrac{s^2b^2}{2} \lim \textstyle\sum_{K_2} \tfrac{g^2(n)}{n} p_k e^{-np_k} + \lim \textstyle\sum_{K_2} g^2(n) p_k e^{-np_k})) \\
&= e^{-\frac{s^2}{2}\left(a^2(c_1+c_2)+b^2(d_1+d_2)\right)}
\end{aligned} \tag{17}$$

which is the characteristic function of a normal distribution with mean zero and variance $a^2(c_1 + c_2) + b^2(d_1 + d_2)$.

Lemma 11. Let $g(n)$ be as in (5). Under Condition 2.3.1,

$$\begin{aligned}
g(n)Z_1 &\xrightarrow{d} N(0, c_1 + c_2), \\
g(n)Z_2 &\xrightarrow{d} N(0, d_1 + d_2).
\end{aligned} \tag{18}$$

Proof. Refer to Zhang and Huang (2008).

Theorem 4. Let $g(n)$ be as in (5). Under Condition 2.3.1,

$$g(n) \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} c_1 + c_2 & 0 \\ 0 & d_1 + d_2 \end{pmatrix} \right).$$

Now, two dimensional asymptotic normality of sample coverage estimators based on Turing's formula was derived. It provides us with one more degree of freedom in distribution

compared with one dimensional case which will be greatly helpful to deal with the case with more than one parameter, such as Hill's model, need to be estimated.

CHAPTER 3: A SUFFICIENT CONDITION FOR THE SECOND ORDER TURING'S FORMULA

3.1 Motivation

Asymptotic normality of Turing's formula has been discussed by Esty (1983) and Zhang and Huang (2008) in different multinomial distribution with infinite categories settings. They both focused on one of the major formula, which is used to estimate the total population proportion of species that are not represented in a sample, in a series of Turing's formulae. For simplicity, let's call it the first order Turing's formula T_1. However it is also useful to discuss the asymptotic properties of another member of Turing's formulae which focus on the estimation of the total population proportion of categories which only contains one sample point. Let's call it the second order Turing's formula T_2. Since both Turing's formula T_1 and T_2 describe the tail behavior of the probability distribution from different aspects. The study on Turing's formula T_2 will help us to acquire more information on the tail behavior besides the Turing's formula T_1. So in this chapter, we derived the asymptotic properties for Turing's formula T_2.

Consider a multinomial distribution with its countably infinite number of categories indexed by $K = \{k; k = 1, \cdots\}$ and its category probabilities denoted by $\{p_k\}$, satisfying $0 < p_k < 1$ for all k and $\sum_{k=1}^{\infty} p_k = 1$. In the subsequent text, the convention that $\sum = \sum_{k=1}^{\infty}$, $\prod = \prod_{k=1}^{\infty}$, $\lim = \lim_{n\to\infty}$ and that $\int = \int_{-\infty}^{+\infty}$, unless otherwise indicated is observed. The symbol "$\sim$" is also used to indicate equality in the limit. Let the category counts in an *iid* sample of size n from the underlying population be denoted by $\{X_k; k \geq 1\}$ and its observed values by $\{x_k; k \geq 1\}$. For a given sample, there are at most n non-zero x_k's. Let, for every integer s, $1 \leq s \leq n$,

$$N_s = \sum 1_{[X_k=s]}, \quad T_s = \binom{n}{s-1}\binom{n}{s}^{-1} N_s, \quad \textit{and} \quad \pi_{s-1} = \sum p_k 1_{[X_k=s-1]}.$$

N_s and π_{s-1} may be thought of as, respectively, the number of categories in the population

that are represented exactly s times in the sample and the total probability associated with all the categories that are represented exactly $s-1$ times in the sample. T_s may be thought of as an estimator of π_{s-1}.

Consider $s = 2$,

$$T_2 = \frac{2N_2}{n-1} \qquad and \qquad \pi_1 = \sum p_k 1_{[X_k=1]}.$$

The objective is to show that, under certain conditions, for some $g(n) > 0$,

$$g(n)(\pi_1 - T_2) \xrightarrow{d} N(0, \sigma^2)$$

where σ^2 is a function of $\{p_k\}$.

3.2 Preliminary Results

Let $K_1 = \{1\}$ and $K_2 = \{2, 3, \cdots\}$. For any $k \in K = K_1 \cup K_2$, let

$$f_k(x) = \begin{cases} p_k & x = 1, \\ -2/(n-1) & x = 2, \\ 0 & x = 0 \text{ or } x \geq 3. \end{cases}$$

$Z = \sum f_k(X_k)$. We are interested in the asymptotic behavior of $Zg(n)$, where $g(n)$ is a function of n satisfying

$$g(n) = O(n^{1-2\delta}) \tag{1}$$

for some $\delta \in (0, 1/4)$, in terms of the limit of its characteristic function, $E[\exp(isZg(n))]$. To begin, we note that $Z = Z_1 + Z_2$, where $Z_1 = \sum_{K_1} f_k(X_k)$ and $Z_2 = \sum_{K_2} f_k(X_k)$. Lemma 12 below is a well-known fact and Lemma 13 is due to Bartlett (1938).

Lemma 12. Let $\{X_k\}$ be the counts of observations in category k, $k = 1, 2, \cdots$, in an *iid* sample under the multinomial model with probability distribution $\{p_k\}$. Then

$$P(X_k = x_k; k = 1, \cdots) = P(Y_k = x_k; k = 1, \cdots | \sum Y_k = n)$$

where $\{Y_k\}$ are independent Poisson random variables with mean np_k.

Lemma 13. Let (U, V) be a two-dimensional random vector with U integer valued. Then

$$E(\exp(ivV|U=n)) = (2\pi P(U=n))^{-1} \int_{-\pi}^{\pi} E[\exp(iu(U-n)+ivV)]du.$$

Thus $E(\exp(isZg(n)))$ is

$$(2\pi P(\sum Y_k = n))^{-1} \int_{-\pi}^{\pi} E[\exp(iu\sum(Y_k - np_k) + isZg(n))]du.$$

We want to evaluate $\lim E(\exp(isZg(n)))$. Toward that end, we first note that, by Stirling's formula, $(2\pi n)^{1/2} P(\sum Y_k = n) \to 1$. Therefore we need only to evaluate the limit of

$$H_n(s) = \frac{\sqrt{n}}{\sqrt{2\pi}} \int_{-\pi}^{\pi} E[\exp(iu\sum(Y_k - np_k) + isZg(n))]du,$$

or letting $t = un^{1/2}$,

$$H_n(s) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} 1[|t| < \pi\sqrt{n}]E[\exp(i(n)^{-1/2}t\sum(Y_k - np_k) + isZg(n))]dt. \qquad (2)$$

Our first task is to allow the limit operator and integral operator to be exchangeable. The key element to support this exchange is (4).

Let

$$\begin{aligned} h_n &= 1[|t| < \pi\sqrt{n}]E[\exp(i(n)^{-1/2}t\textstyle\sum(Y_k - np_k) + isZg(n)] \\ h_{n1} &= 1[|t| < \pi\sqrt{n}]E[\exp(i(n)^{-1/2}t(Y_1 - np_1) + isZ_1 g(n)] \qquad (3) \\ h_{n2} &= 1[|t| < \pi\sqrt{n}]E[\exp(i(n)^{-1/2}t\textstyle\sum_{K_2}(Y_k - np_k) + isZ_2 g(n)], \end{aligned}$$

$$H_n(s) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} h_n dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} h_{n1} h_{n2} dt.$$

Since $|h_{n2}| \le 1$, $|h_n| \le |h_{n1}|$. On the other hand,

$$
\begin{aligned}
&E[\exp(iu(Y_1 - np_1) + isf_1(Y_1)g(n))] \\
&= \exp(iu(-np_1) + 0)\exp(-np_1) + \exp(iu(1 - np_1) + isp_1 g(n))np_1 \exp(-np_1) \\
&\quad + \exp(iu(2 - np_1) - is\tfrac{2}{n-1}g(n))\tfrac{(np_1)^2}{2}\exp(-np_1) \\
&\quad + \textstyle\sum_{j=3}^{\infty}\exp(iu(j - np_1))P(Y_1 = j) \\
&= \textstyle\sum_{j=0}^{\infty}\exp(iu(j - np_1))P(Y_1 = j) \\
&\quad - \exp(iu(1 - np_1))np_1\exp(-np_1) - \exp(iu(2 - np_1))\tfrac{(np_1)^2}{2}\exp(-np_1) \\
&\quad + \exp(iu(1 - np_1) + isp_1 g(n))np_1\exp(-np_1) + \exp(iu(2 - np_1) - is\tfrac{2}{n-1}g(n))\tfrac{(np_1)^2}{2}\exp(-np_1) \\
&= [\exp(-iunp_1)\exp(i\sin(u)np_1)\exp(np_1(\cos(u) - 1))] \\
&\quad - \exp(iu(1 - np_1))np_1\exp(-np_1) - \exp(iu(2 - np_1))\tfrac{(np_1)^2}{2}\exp(-np_1) \\
&\quad + \exp(iu(1 - np_1) + isp_1 g(n))np_1\exp(-np_1) + \exp(iu(2 - np_1) - is\tfrac{2}{n-1}g(n))\tfrac{(np_1)^2}{2}\exp(-np_1).
\end{aligned}
$$

Therefore (recall $t = u\sqrt{n}$),

$$
|h_{n1}| \le 1[|t| < \pi\sqrt{n}]\left[\exp(np_1(\cos(tn^{-\frac{1}{2}}) - 1)) + 2[np_1\exp(-np_1) + \frac{(np_1)^2}{2}\exp(-np_1)]\right] \quad (= \overline{h}_{n1}).
$$

It is clear that, for any t, by Taylor's formula for $\cos(x)$,

$$
\lim \overline{h}_{n1} = \lim 1[|t| < \pi\sqrt{n}]\exp(np_1(\cos(tn^{-1/2}) - 1)) = \exp(-p_1 t^2/2) \qquad (= \overline{h}_1).
$$

$$\begin{aligned}
\int |\overline{h}_{n1}|dt &= \int 1[|t| < \pi\sqrt{n}] \left[\exp(np_1(\cos(tn^{-\frac{1}{2}}) - 1))\right] dt \\
&\quad +2\int 1[|t| < \pi\sqrt{n}]np_1 \exp(-np_1)dt + 2\int 1[|t| < \pi\sqrt{n}]\tfrac{(np_1)^2}{2} \exp(-np_1)dt \\
&= \int 1[|t| < \pi\sqrt{n}] \left[\exp(np_1(\cos(tn^{-\frac{1}{2}}) - 1))\right] dt \\
&\quad +2 \times 2\pi\sqrt{n}np_1 \exp(-np_1) + 2 \times 2\pi\sqrt{n}\tfrac{(np_1)^2}{2} \exp(-np_1)
\end{aligned}$$

Since the last two terms above vanish to zero as $n \to \infty$, we have, letting θ be a constant in $(0, 1/2)$,

$$\begin{aligned}
\lim \int |\overline{h}_{n1}|dt &= \lim \int 1[|t| < \pi\sqrt{n}] \left[\exp(np_1(\cos(tn^{-\frac{1}{2}}) - 1))\right] dt \\
&= \lim \textstyle\int_{-\pi}^{+\pi} \sqrt{n} \left[\exp(np_1(\cos(u) - 1))\right] du \\
&= \lim \textstyle\int_{|u|<\frac{1}{n^{(1-\theta)/2}}} \sqrt{n} \left[\exp(np_1(\cos(u) - 1))\right] du \\
&\quad + \lim \textstyle\int_{\frac{1}{n^{(1-\theta)/2}} \le |u| < \pi} \sqrt{n} \left[\exp(np_1(\cos(u) - 1))\right] du \\
&(= \lim \eta_1 + \lim \eta_2).
\end{aligned}$$

The second term of the last expression above is zero. To see this, we note that for any u

satisfying $\frac{1}{n^{(1-\theta)/2}} \leq |u| < \pi$, $\cos(u) - 1 \leq \cos(1/n^{(1-\theta)/2}) - 1$, and hence

$$\begin{aligned}
\lim \eta_2 &\leq \lim \int_{\frac{1}{n^{(1-\theta)/2}} \leq |u| < \pi} \sqrt{n} \left[\exp(np_1(\cos(1/n^{(1-\theta)/2}) - 1))\right] du \\
&\leq \lim 2\pi\sqrt{n} \left[\exp(np_1(\cos(1/n^{(1-\theta)/2}) - 1))\right] \\
&= \lim 2\pi\sqrt{n} \left[\exp\left(-np_1\left(1 - \cos(1/n^{(1-\theta)/2})\right)\right)\right] \\
&= \lim 2\pi\sqrt{n} \left[\exp\left(-np_1\left(\frac{\sin^2(1/n^{(1-\theta)/2})}{1+\cos(1/n^{(1-\theta)/2})}\right)\right)\right] \\
&= \lim 2\pi\sqrt{n} \exp\left(-np_1 O(\tfrac{1}{n^{1-\theta}})\right) \\
&= \lim 2\pi\sqrt{n} \exp\left(-p_1 O(n^{\theta})\right) = 0.
\end{aligned}$$

For u satisfying $|u| < \frac{1}{n^{(1-\theta)/2}}$, consider the Taylor expansion of

$$\begin{aligned}
\cos(u) - 1 &= -\tfrac{u^2}{2!} + \tfrac{u^4}{4!} - \tfrac{u^6}{6!} + \cdots + \tfrac{(-1)^m u^{2m}}{(2m)!} + \cdots \\
&\leq -\tfrac{u^2}{2} + (u^4 + u^8 + \cdots + u^{4m} + \cdots) \\
&= -\tfrac{u^2}{2} + \tfrac{u^4}{1-u^4}.
\end{aligned}$$

Therefore

$$\begin{aligned}
\lim \eta_1 &\leq \lim \textstyle\int_{|u|<\frac{1}{n^{(1-\theta)/2}}} \sqrt{n} \exp\left(np_1 \left(-\frac{u^2}{2} + \frac{\frac{1}{n^{2-2\theta}}}{1-\frac{1}{n^{2-2\theta}}} \right) \right) du \\
&= \lim \textstyle\int_{|u|<\frac{1}{n^{(1-\theta)/2}}} \sqrt{n} \exp\left(-\frac{np_1u^2}{2} + np_1 \frac{\frac{1}{n^{2-2\theta}}}{1-\frac{1}{n^{2-2\theta}}} \right) du \\
&= \lim \left[\left(\textstyle\int_{|u|<\frac{1}{n^{(1-\theta)/2}}} \sqrt{n} \exp\left(-\frac{np_1u^2}{2} \right) du \right) \exp\left(O(\frac{1}{n^{1-2\theta}}) \right) \right] \\
&\quad (\text{ letting } t = u\sqrt{n}) \\
&= \lim \left[\left(\textstyle\int_{|t|<n^{\theta/2}} \exp\left(-\frac{p_1t^2}{2} \right) dt \right) \exp\left(O(\frac{1}{n^{1-2\theta}}) \right) \right] \\
&= \textstyle\int \exp\left(-p_1t^2/2 \right) dt.
\end{aligned}$$

Since $\cos(u) - 1 \geq -\frac{u^2}{2}$ for all u satisfying $|u| < \frac{1}{n^{(1-\theta)/2}}$, it is easy to establish $\lim \eta_1 \geq \int \exp\left(-p_1t^2/2 \right) dt$, and hence $\lim \eta_1 = \int \exp\left(-p_1t^2/2 \right) dt$.

Now that we have established

$$\lim \int |\overline{h}_{n1}| dt = \int \lim |\overline{h}_{n1}| dt, \tag{4}$$

by the Dominated Convergence Theorem, we have the following lemma.

Lemma 14. Let h_n and H_n be as defined in (2) and (3) respectively. Then

$$\lim H_n = \frac{1}{\sqrt{2\pi}} \int \lim h_n dt.$$

Let,

$$h_n(s) = 1[|t| < \pi\sqrt{n}] E[\exp(i(n)^{-1/2} t \textstyle\sum (Y_k - np_k) + isZg(n)]$$

For each k, it can be verified that, letting

$$B_k = \exp(-itp_k n^{1/2})[\exp(np_k(\exp(itn^{-1/2}) - 1))]$$

$$C_k = \exp(-itp_k n^{1/2}) \exp(itn^{-1/2})[\exp(isp_k g(n)) - 1]np_k \exp(-np_k)$$

$$D_k = \exp(-itp_k n^{1/2}) \exp(2itn^{-1/2})[\exp(-is\tfrac{2}{n-1}g(n)) - 1]\tfrac{(np_k)^2}{2} \exp(-np_k),$$

and $E_k = C_k + D_k$, $h_n \sim \prod(B_k + E_k)$. We are interested in evaluating $\lim \prod(B_k + E_k)$.

The facts of the following two lemmas are given by Esty (1983).

Lemma 15. Let $\{\beta_k\}$ and $\{\epsilon_k\}$ be two sequences of complex numbers, and M_n be a sequence of subsets of K, indexed by n. If

1. $\prod_{M_n} \beta_k \sim \beta$,
2. $(\sum_{M_n} \epsilon_k) \sim \epsilon$,
3. $\beta_k \sim 1$ uniformly,
4. $\epsilon_k \sim 0$ uniformly,
5. there exists a constants, δ_1 such that, $\sum_{M_n} |\beta_k - 1| \leq \delta_1$, and
6. there exists a constants, δ_2 such that, $\sum_{M_n} |\epsilon_k| \leq \delta_2$;

then

$$\prod_{M_n} (\beta_k + \epsilon_k) \sim \beta e^{\epsilon}$$

where β and ϵ may also depend on n.

Lemma 16. For all $k \in K$,

$$B_k = \exp[(-t^2/2)p_k + O(t^3 p_k n^{-1/2})].$$

The next lemma includes three useful facts.

Lemma 17. 1. For any complex number x satisfying $|x| < 1$, $|\ln(1+x)| \leq \frac{|x|}{1-|x|}$.

2. For any real number $x \in [0,1)$, $1 - x \geq \exp\left(-\frac{x}{1-x}\right)$.

3. For any real number $x \in (0, 1/2)$, $\frac{1}{1-x} < 1 + 2x$.

Proof. 1) By Taylor's formula, $|\ln(1+x)| = |\sum_{j=1}^{\infty}(-1)^{j+1}x^j/j| \leq \sum_{j=1}^{\infty}|x|^j = |x|/(1-|x|)$.

2) The function $y = \frac{1}{1+t}e^t$ is strictly increasing over $[0,\infty)$, and has value 1 at $t = 0$. Therefore $\frac{1}{1+t}e^t \geq 1$ for $t \in [0,\infty)$. The desired inequality follows the change of variable $x = t/(1+t)$.

3) The proof is trivial.

Let us consider a partition of the index set $K = I \cup II$ where

$$I = \{k; p_k \leq \frac{\sqrt{2}}{n^{1-\delta}}\} \quad \text{and} \quad II = \{k; p_k > \frac{\sqrt{2}}{n^{1-\delta}}\}$$

where δ is as in (1).

Lemma 18. (a) $\sum_{II}|E_k| \to 0$; and (b) $\prod_{II}(B_k + E_k)/\prod_{II}B_k \to 1$.

Proof. (a) $\sum_{II}|E_k| \leq 2\sum_{II}(np_ke^{-np_k} + \frac{(np_k)^2}{2}e^{-np_k})$. Since the derivative of $(np_ke^{-np_k} + \frac{(np_k)^2}{2}e^{-np_k})$ for any $k \in II$, with respect to p_k, is negative. $(np_ke^{-np_k} + \frac{(np_k)^2}{2}e^{-np_k})$ attains its maximum at $\frac{\sqrt{2}}{n^{1-\delta}}$, with value $\sqrt{2}n^{\delta}e^{-\sqrt{2}n^{\delta}} + n^{2\delta}e^{-\sqrt{2}n^{\delta}}$. The total number of indices in II is less or equal to $\frac{n^{1-\delta}}{\sqrt{2}}$. Therefore

$$\sum_{II}|E_k| \leq 2(\frac{n^{1-\delta}}{\sqrt{2}})(\sqrt{2}n^{\delta}e^{-\sqrt{2}n^{\delta}} + n^{2\delta}e^{-\sqrt{2}n^{\delta}}) = \sqrt{2}e^{-\sqrt{2}n^{\delta}}(\sqrt{2}n + n^{1+\delta}) \to 0. \qquad (5)$$

(b) By Lemma 16, $|B_k|$ is bounded away from zero, and by the fact that $\lim|E_k| = 0$ (and hence $\lim|E_k|/|B_k| = 0$), and by applying the first part of Lemma 17 with $x = E_k/B_k$,

we have

$$|\ln[\textstyle\prod_{II}(B_k+E_k)/\prod_{II}B_k]| = \left|\textstyle\sum_{II}\ln\left(1+\frac{E_k}{B_k}\right)\right| \le \textstyle\sum_{II}\left|\ln\left(1+\frac{E_k}{B_k}\right)\right|$$

$$\le \textstyle\sum_{II}\left(\frac{|E_k|}{|B_k|-|E_k|}\right) = O(\textstyle\sum_{II}|E_k|) \to 0.$$

Now let us state the condition under which many of the subsequent results are established.

Condition 3.2.1. As $n \to \infty$,

1. $\sum g^2(n)p_k^2 e^{-np_k} \to c_2 \ge 0$,
2. $\sum g^2(n)np_k^3 e^{-np_k} \to c_3 \ge 0$ and
3. $c_2 + c_3 > 0$.

Lemma 19. Under Condition 3.2.1, all the conditions of Lemma 15 are satisfied with $M_n = I$, $\beta_k = B_k$, $\beta = B$, $\epsilon_k = E_k$, and $\epsilon = E$.

Proof. We need to check all six conditions in Lemma 15.

3) is true because

$$B_k = \exp(-(t^2/2)p_k)\exp(O((t^3/\sqrt{n})p_k))),$$

and p_k and $p_k/\sqrt{n}$ are uniformly bounded by $\frac{\sqrt{2}}{n^{1-\delta}}$ and $\frac{\sqrt{2}}{n^{1-\delta}\sqrt{n}}$ respectively for $k \in I$.

For 1), since $\sum_I p_k \to 0$,

$$\prod_I B_k = \exp(-(t^2/2)\sum_I p_k)\exp(O((t^3/\sqrt{n})\sum_I p_k))) \to 1.$$

For 2), 4) and 6),

$$E_k = np_k e^{-np_k} e^{-itp_k\sqrt{n}} \left\{ \left[1 + \frac{it}{\sqrt{n}} - \frac{t^2}{2n} + O(\frac{t^3}{n^{3/2}}) \right] \left[isp_k g(n) - \frac{s^2 p_k^2 g^2(n)}{2} + O(s^3 g^3(n) p_k^3) \right] \right.$$

$$\left. + \frac{np_k}{2} \left[1 + \frac{2it}{\sqrt{n}} - \frac{4t^2}{2n} + O(\frac{t^3}{n^{3/2}}) \right] \left[-\frac{2isg(n)}{n-1} - \frac{4s^2 g^2(n)}{(n-1)^2} + O(\frac{s^3 g^3(n)}{(n-1)^3}) \right] \right\}$$

$$= np_k e^{-np_k} e^{-itp_k\sqrt{n}} \Big\{ isp_k g(n) - \tfrac{stp_k g(n)}{\sqrt{n}} - \tfrac{ist^2 p_k g(n)}{2n} + O(\tfrac{ist^3 p_k g(n)}{n^{3/2}})$$

$$- \tfrac{s^2 p_k^2 g^2(n)}{2} - \tfrac{its^2 p_k^2 g^2(n)}{2\sqrt{n}} + \tfrac{s^2 t^2 p_k^2 g^2(n)}{4n} + O(\tfrac{s^2 t^3 p_k^2 g^2(n)}{2n^{3/2}})$$

$$+ O(s^3 p_k^3 g^3(n)) + \tfrac{it}{\sqrt{n}} O(s^3 p_k^3 g^3(n)) - \tfrac{t^2}{2n} O(s^3 p_k^3 g^3(n)) + O(\tfrac{t^3}{n^{3/2}}) O(s^3 p_k^3 g^3(n))$$

$$- \tfrac{isnp_k g(n)}{n-1} + \tfrac{2tnp_k sg(n)}{\sqrt{n}(n-1)} + \tfrac{2isp_k t^2 g(n)}{n-1} + O(-\tfrac{2isg(n)t^3 p_k}{2\sqrt{n}(n-1)})$$

$$- \tfrac{4np_k s^2 g^2(n)}{2(n-1)^2} - \tfrac{4itnp_k s^2 g^2(n)}{\sqrt{n}(n-1)^2} + \tfrac{4p_k t^2 s^2 g^2(n)}{(n-1)^2} + O(\tfrac{4s^2 g^2(n) t^3 p_k}{2\sqrt{n}(n-1)^2})$$

$$+ \tfrac{np_k}{2} O(\tfrac{s^3 g^3(n)}{(n-1)^3}) + \tfrac{itnp_k}{\sqrt{n}} O(\tfrac{s^3 g^3(n)}{(n-1)^3}) - p_k t^2 O(\tfrac{s^3 g^3(n)}{(n-1)^3}) + O(\tfrac{t^3 p_k}{2\sqrt{n}}) O(\tfrac{s^3 g^3(n)}{(n-1)^3}) \Big\}$$

$$= np_k e^{-np_k} e^{-itp_k\sqrt{n}} \Big\{ \Big(isp_k g(n) - \tfrac{isnp_k g(n)}{n-1} \Big) - \tfrac{stp_k g(n)}{\sqrt{n}} - \tfrac{ist^2 p_k g(n)}{2n}$$

$$- \tfrac{s^2 p_k^2 g^2(n)}{2} - \tfrac{its^2 p_k^2 g^2(n)}{2\sqrt{n}} + \tfrac{s^2 t^2 p_k^2 g^2(n)}{4n}$$

$$+ \tfrac{2t\sqrt{n} p_k sg(n)}{n-1} + \tfrac{2isp_k t^2 g(n)}{n-1} - \tfrac{2np_k s^2 g^2(n)}{(n-1)^2}$$

$$- \tfrac{4it\sqrt{n} p_k s^2 g^2(n)}{(n-1)^2} + \tfrac{4p_k t^2 s^2 g^2(n)}{(n-1)^2}$$

$$+ O(\tfrac{ist^3 p_k g(n)}{n^{3/2}}) + O(\tfrac{s^2 t^3 p_k^2 g^2(n)}{2n^{3/2}}) + O(s^3 p_k^3 g^3(n))$$

$$+ O(\tfrac{is^3 tp_k^3 g^3(n)}{\sqrt{n}}) - O(\tfrac{s^3 t^2 p_k^3 g^3(n)}{2n}) + O(\tfrac{s^3 t^3 p_k^3 g^3(n)}{n^{3/2}})$$

$$+ O(\tfrac{ist^3 p_k g(n)}{\sqrt{n}(n-1)}) + O(\tfrac{2s^2 t^3 p_k g^2(n)}{\sqrt{n}(n-1)^2}) + O(\tfrac{ns^3 p_k g^3(n)}{2(n-1)^3})$$

$$+ O(\tfrac{is^3 t\sqrt{n} p_k g^3(n)}{(n-1)^3}) - O(\tfrac{s^3 t^2 p_k g^3(n)}{(n-1)^3}) + O(\tfrac{s^3 t^3 p_k g^3(n)}{2\sqrt{n}(n-1)^3}) \Big\}$$

(6)

Now we observe the following:

1. For all $k \in I$, $\exp(-itp_k\sqrt{n}) \to 1$ uniformly since $p_k\sqrt{n} \leq \frac{\sqrt{2}\sqrt{n}}{n^{1-\delta}} \to 0$.

2. It is easily checked that every additive term of E_k converges to zero uniformly for all $k \in I$. Therefore 4) is checked.

3. It is easily checked that, for every term within the curly brackets in (6), denoted by $\tau(s,t,n,p_k)$, except the fourth and ninth terms,

$$\sum_I e^{-np_k}|\tau(s,t,n,p_k)| \leq \sum e^{-np_k}|\tau(s,t,n,p_k)| \to 0$$

uniformly by Condition 3.2.1.

The uniform convergence of $\sum_I np_k^3 g^2(n)e^{-np_k}$ and $\sum_I \frac{n^2}{(n-1)^2}p_k^2 g^2(n)e^{-np_k}$ are directly guaranteed by Condition 3.2.1. Therefore 2) is checked. The uniformity of the convergence for $\sum_I E_k$ and hence for $\sum_I |E_k|$ guarantees 6).

For 5), since $B_k = \exp\left(-\frac{t^2}{2}p_k + O(t^3p_kn^{-1/2})\right)$ and $-\frac{t^2}{2}p_k + O(t^3p_kn^{-1/2}) \to 0$ uniformly, we have

$$|B_k - 1| \leq \frac{|-\frac{t^2}{2}p_k + O(t^3p_kn^{-1/2})|}{1 - |-\frac{t^2}{2}p_k + O(t^3p_kn^{-1/2})|} = O\left(-\frac{t^2}{2}p_k + t^3p_kn^{-1/2}\right)$$

and hence

$$\sum_I |B_k - 1| \leq O\left(\frac{t^2}{2}\sum_I p_k) + \frac{|t^3|}{\sqrt{n}}\sum_I p_k\right) < O(t^2 + |t^3|).$$

Lemma 15 and Lemma 19 give immediately the following corollary.

Corollary 4. Under Condition 3.2.1, $\prod_I (B_k + E_k) \sim \prod_I B_k \exp(\sum_I E_k)$.

Lemma 20. Under Condition 3.2.1, $\prod(B_k + E_k) \to Be^E$, where $B = \lim \prod B_k$ and $E = \lim \sum E_k$.

Proof.

$$\begin{aligned}\prod(B_k+E_k) &= \prod_I(B_k+E_k)\prod_{II}(B_k+E_k) \sim \prod_I(B_k+E_k)\prod_{II}B_k \qquad \text{(by Lemma 18)}\\ &\sim \prod_I B_k(\exp\textstyle\sum_I E_k)\prod_{II}B_k \qquad \text{(by Lemma 19)}\\ &\sim \prod B_k(\exp\textstyle\sum E_k) \qquad \text{(by Lemma 18).}\end{aligned}$$

Remark 1. At this point, one may see the reason why it is imposed that $g(n) = O(n^{1-2\delta})$ for some small positive δ. If $g(n)$ is let to be a sequence increasing to infinity in the order of n or faster, $\sum_{II} E_k \to 0$ cannot be established using the current method. The proof for (a) of Lemma 18 will break down. Consequently, the partition of $K = I \cup II$ will not effectively support the subsequent proofs.

3.3 Main Results

Theorem 5. Let $g(n)$ be as in (1). Under Condition 3.2.1,

$$g(n)(\pi_1 - T_2) \xrightarrow{d} N(0, 4c_2 + c_3).$$

Proof. Since $\lim\prod B_k = e^{-\frac{t^2}{2}}$ and

$$\lim\sum E_k = -\frac{s^2}{2}\left(\lim\sum np_k^3 g^2(n)e^{-np_k} + \lim\sum 4\frac{n^2}{(n-1)^2}p_k^2 g^2(n)e^{-np_k}\right),$$

$$\lim H_n = \left(\frac{1}{\sqrt{2\pi}}\int e^{-\frac{t^2}{2}}dt\right)e^{-\frac{s^2}{2}\left(\lim\sum np_k^3 g^2(n)e^{-np_k} + \lim\sum 4\frac{n^2}{(n-1)^2}p_k^2 g^2(n)e^{-np_k}\right)} = e^{-\frac{s^2}{2}(c_3+4c_2)}$$

which is the characteristic function of a normal distribution with mean zero and variance $c_3 + 4c_2$.

Given a $g(n)$ satisfying (1), Condition 3.2.1 imposes a rate of convergence for $\{p_k\}$. To see that and that the condition of Theorem 5 describes a non-empty class of distribution,

we consider the following example.

Example 3.3.1. Let $p_k = \frac{c}{(k+1)^2}$, $k = 1, \cdots$, where $c = \frac{1}{(\pi^2/6)-1}$. Then $g(n)$ must be of order $O(n^{3/4})$ for Condition 3.2.1 to hold.

To see this, we have

$$\begin{aligned} g^2(n)\int_1^\infty \frac{c^2}{(x+1)^4} e^{-\frac{cn}{(x+1)^2}}\,dx &= c^2 g^2(n)\int_0^{1/2} t^2 e^{-cnt^2}\,dt \\ &= c^2 g^2(n)\int_0^{\frac{\sqrt{2cn}}{2}} \frac{t^2}{2cn} e^{-\frac{t^2}{2}} \frac{1}{\sqrt{2cn}}\,dt \qquad (7) \\ &= O(\tfrac{g^2(n)}{n\sqrt{n}}). \end{aligned}$$

The last expression goes to a non-zero constant if and only if $g(n) = O(n^{3/4})$.

Similarly,

$$\begin{aligned} g^2(n)\int_1^\infty n\frac{c^3}{(x+1)^6} e^{-\frac{cn}{(x+1)^2}}\,dx &= c^3 g^2(n) n\int_1^\infty \frac{1}{(x+1)^6} e^{-\frac{cn}{(x+1)^2}}\,dx \\ &= c^3 g^2(n) n\int_0^{\frac{1}{2}} t^6 e^{-cnt^2} \frac{1}{t^2}\,dt \\ &= c^3 g^2(n) n\int_0^{\frac{1}{2}} t^4 e^{-cnt^2}\,dt \qquad (8) \\ &= c^3 g^2(n) n\int_0^{\frac{\sqrt{2cn}}{2}} \frac{t^4}{4c^2n^2} e^{-\frac{t^2}{2}} \frac{1}{\sqrt{2cn}}\,dt \\ &= O(\tfrac{g^2(n)}{n\sqrt{n}}). \end{aligned}$$

The last expression goes to a non-zero constant if and only if $g(n) = O(n^{3/4})$.

Let us consider the following condition:

Condition 3.3.1. As $n \to \infty$,

1. $\frac{g^2(n)}{n^2} E(N_2) \to \frac{c_2}{2} \geq 0$,
2. $\frac{g^2(n)}{n^2} E(N_3) \to \frac{c_3}{6} \geq 0$, and

3. $c_2 + c_3 > 0$.

Lemma 21. Condition 3.2.1 and Condition 3.3.1 are equivalent.

Proof. Let us again consider the partition of $K = I \cup II$. First we note that $p^2 e^{-np}$ has a negative derivative with respect to p on interval $(2/n, 1]$ and hence on $(\sqrt{2}/n^{1-\delta}, 1]$ for large n. Therefore, since there are at most $\frac{n^{1-\delta}}{\sqrt{2}}$ terms in II,

$$0 \leq \tfrac{g^2(n)}{n^2} C_n^2 \textstyle\sum_{II} p_k^2 (1-p_k)^{n-2} \leq \tfrac{g^2(n)}{n^2} C_n^2 \sum_{II} p_k^2 e^{-(n-2)p_k} \leq \tfrac{g^2(n)}{n^2} C_n^2 \sum_{II} (\tfrac{\sqrt{2}}{n^{1-\delta}})^2 e^{-\frac{(n-2)\sqrt{2}}{n^{1-\delta}}}$$

$$\leq \tfrac{g^2(n)}{n^2} C_n^2 \tfrac{n^{1-\delta}}{\sqrt{2}} (\tfrac{\sqrt{2}}{n^{1-\delta}})^2 e^{-\frac{(n-2)\sqrt{2}}{n^{1-\delta}}} = O(n^{1-3\delta}) e^{-O(n^\delta)} \to 0.$$

Thus we have

$$\lim \frac{g^2(n)}{n^2} E(N_2) = \lim \frac{g^2(n)}{n^2} C_n^2 \sum_I p_k^2 (1-p_k)^{n-2} \tag{9}$$

and

$$\lim g^2(n) \sum p_k^2 \exp(-np_k) = \lim g^2(n) \sum_I p_k^2 \exp(-np_k). \tag{10}$$

On the other hand,

$$\tfrac{g^2(n)}{n^2} C_n^2 \textstyle\sum_I p_k^2 (1-p_k)^{n-2} \leq \tfrac{g^2(n)}{n^2} C_n^2 \sum_I p_k^2 e^{-(n-2)p_k} \leq \tfrac{g^2(n)}{n^2} C_n^2 \exp(2 \sup_I p_k) \textstyle\sum_I p_k^2 e^{-np_k}.$$

Furthermore, applying 2) and 3) of Lemma 17 in the first and the third steps below respectively, we have

$$\tfrac{g^2(n)}{n^2} C_n^2 \textstyle\sum_I p_k^2 (1-p_k)^{n-2} \geq \tfrac{g^2(n)}{n^2} C_n^2 \sum_I p_k^2 \exp\left(-\tfrac{(n-2)p_k}{1-p_k}\right) \geq \tfrac{g^2(n)}{n^2} C_n^2 \sum_I p_k^2 \exp\left(-\tfrac{np_k}{1-\sup_I p_k}\right)$$

$$\geq \tfrac{g^2(n)}{n^2} C_n^2 \exp(-2n(\sup_I p_k)^2) \textstyle\sum_I p_k^2 e^{-np_k}.$$

Noting the fact that $\lim \exp(2 \sup_I p_k) = 1$ and $\lim \exp(-2n(\sup_I p_k)^2) = 1$ by the

definition of I,

$$\lim \frac{g^2(n)}{n^2} C_n^2 \sum_I p_k^2 (1-p_k)^{n-2} = \lim \frac{g^2(n)}{n^2} C_n^2 \sum_I p_k^2 e^{-np_k},$$

and hence by (11) and (12), we have the equivalence of the first parts of Condition 3.2.1 and Condition 3.3.1:

$$\lim \frac{g^2(n)}{n^2} E(N_2) = \frac{1}{2} \lim \sum g^2(n) p_k^2 \exp(-np_k).$$

The equivalence of the second parts can be established similarly.

Let us again consider the partition of $K = I \cup II$. First we note that $p^3 e^{-np}$ has a negative derivative with respect to p on interval $(3/n, 1]$ and hence on $(\sqrt{2}/n^{1-\delta}, 1]$ for large n. Therefore, since there are at most $\frac{n^{1-\delta}}{\sqrt{2}}$ terms in II,

$$0 \le \tfrac{g^2(n)}{n^2} C_n^3 \textstyle\sum_{II} p_k^3 (1-p_k)^{n-3} \le \tfrac{g^2(n)}{n^2} C_n^3 \sum_{II} p_k^3 e^{-(n-3)p_k} \le \tfrac{g^2(n)}{n^2} C_n^3 \sum_{II} (\tfrac{\sqrt{2}}{n^{1-\delta}})^3 e^{-\frac{(n-3)\sqrt{2}}{n^{1-\delta}}}$$

$$\le \tfrac{g^2(n)}{n^2} C_n^3 \tfrac{n^{1-\delta}}{\sqrt{2}} (\tfrac{\sqrt{2}}{n^{1-\delta}})^3 e^{-\frac{(n-3)\sqrt{2}}{n^{1-\delta}}} = O(n^{1-2\delta}) e^{-O(n^\delta)} \to 0.$$

Thus we have

$$\lim \frac{g^2(n)}{n^2} E(N_3) = \lim \frac{g^2(n)}{n^2} C_n^3 \sum_I p_k^3 (1-p_k)^{n-3} \tag{11}$$

and

$$\lim g^2(n) n \sum p_k^3 \exp(-np_k) = \lim g^2(n) n \sum_I p_k^3 \exp(-np_k). \tag{12}$$

On the other hand,

$$\tfrac{g^2(n)}{n^2} C_n^3 \textstyle\sum_I p_k^3 (1-p_k)^{n-3} \le \tfrac{g^2(n)}{n^2} C_n^3 \sum_I p_k^3 e^{-(n-3)p_k} \le \tfrac{g^2(n)}{n^2} C_n^3 \exp(3 \sup_I p_k) \sum_I p_k^3 e^{-np_k}.$$

Furthermore, applying 2) and 3) of Lemma 17 in the first and the third steps below

respectively, we have

$$\frac{g^2(n)}{n^2}C_n^3\sum_I p_k^3(1-p_k)^{n-3} \geq \frac{g^2(n)}{n^2}C_n^3\sum_I p_k^3\exp\left(-\frac{(n-3)p_k}{1-p_k}\right) \geq \frac{g^2(n)}{n^2}C_n^3\sum_I p_k^3\exp\left(-\frac{np_k}{1-\sup_I p_k}\right)$$

$$\geq \frac{g^2(n)}{n^2}C_n^3\exp(-2n(\sup_I p_k)^2)\sum_I p_k^3 e^{-np_k}.$$

Noting the fact that $\lim\exp(3\sup_I p_k) = 1$ and $\lim\exp(-2n(\sup_I p_k)^2) = 1$ by the definition of I,

$$\lim\frac{g^2(n)}{n^2}C_n^3\sum_I p_k^3(1-p_k)^{n-3} = \lim\frac{g^2(n)}{n^2}C_n^3\sum_I p_k^3 e^{-np_k},$$

and hence by (11) and (12), we have the equivalence of the first parts of Condition 3.2.1 and Condition 3.3.1:

$$\lim\frac{g^2(n)}{n^2}E(N_3) = \frac{1}{6}\lim\sum g^2(n)np_k^3\exp(-np_k).$$

Lemma 21 allows us to re-state Theorem 5:

Theorem 6. If there exists a $g(n)$ satisfying (1) and Condition 3.3.1, then

$$\frac{n(\pi_1 - T_2)}{\sqrt{8E(N_2) + 6E(N_3)}} \xrightarrow{d} N(0,1). \tag{13}$$

As a consequence of Theorem 5, we have the following theorem:

Theorem 7. If there exists a $g(n)$ satisfying (1) and Condition 3.3.1, then

$$\frac{n(\pi_1 - T_2)}{\sqrt{8N_2 + 6N_3}} \xrightarrow{d} N(0,1).$$

The proof of Theorem 7 is similar as in Zhang and Huang (2008). (omitted)

We note that the conditions of Theorems 6 and 7 requires no further knowledge of $g(n)$ other than its existence.

Theorem 7 leads to an approximate $(1-\alpha)$-level confidence interval for π_1:

$$T_2 \pm z_{\alpha/2}\sqrt{8N_2/n^2 + 6N_3/n^2}. \tag{14}$$

REFERENCES

Bartlett, M.S. (1938), *The characteristic function of a conditional statistic*, Journal of the London Mathematical Society, 13, pp.62-67.

Chao, A. (1981), *On estimating the probability of discovering a new species*, Annals of Statistics, 9, pp.1339-1342.

Chao, A. (1984), *Nonparametric estimation of the number of the classes in a population*, Scandinavian Journal of Statistics, 11, pp.265-270.

Chao, A. and Lee, S. (1992), *Estimating the number of classes via sample covergae*, Journal of American Statistical Assciation, 87, pp.210-217.

Efron, B. and Thisted, R. (1976), *Estimating the number of unseen species: how many words did Shakespeare know?*, Biometrika, 63, pp.435-447.

Esty, W.W. (1982), *Confidence intervals for the coverage of low coverage samples*, Annals of Statistics, 10, pp.190-196.

Esty, W.W. (1983), *A normal limit law for a nonparametric estimator of the coverage of a random sample*, Annals of Statistics, 11, pp.905-912.

Esty, W.W. (1985), *Estimation of the number of classes in a population and the coverage of a sample*, Mathematical Scientist, 10, pp.41-50.

Esty, W.W. (1986a), *The size of a coinage*, Numismatic Chronicle, 146, pp.185-215.

Esty, W.W. (1986b), *The efficiency of Good's nonparametric coverage estimator*, Annals of Statistics, 14, pp.1257-1260.

Fritsch, K.S. and Hsu, J.C. (1999), *Multiple Comparison of Entropies with Application to Dinosaur Biodiversity*, Biometrics, 55, pp.1300-1305.

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

Good, I.J.and Toulmin, G.H. (1956), *The number of new species, and the increase in population coverage, when a sample is increased*, Biometrika, 43, pp.45-63.

Harris, B. (1959), *Determining bounds on integrals with applications to cataloging problems*, Annals of Mathematical Statististics, 30, pp.521-548.

Harris, B. (1968), *Statistical inference in the classical occupancy problem unbiased estimation of number of classes*, Journal of American Statistical Assciation, 63, pp.837-847.

Hellmann, J.J. and Fowler, G.W. (1999), *Bias, precision, and accuracy of four measures of species richness*, Ecological Applications, 9(3), pp. 824-834.

Hill, M. Bruce (1975), *A simple general approach to inference about the tail of a distribution*, The Annals of Statistics, 3, pp. 1163-1174.

Hill, M.O. (1973), *Diversity and evenness: a unifying notation and its consequences*, Ecology, 54, pp. 427-431.

Holst, L. (1981), *Some assymptotic results for incomplete multinomial or Poisson samples*, Scandinavian Journal of Statistics, 8, pp.243-246.

Magurran, A.E. (1988), *Ecological diversity and its measurement*, Princeton University Press, Princeton New Jersey, USA.

Mao, C.X. and Lindsay, B.G. (2002), *A Poisson model for the coverage problem with a genomic application*, Biometrika, 89, pp.669-681.

Rennolls, K. and Laumonier, Y. (2006), *A new local estimator of regional species diversity, in terms of 'shadow species', with a case study from Sumatra*, Journal of Tropical Ecology, 22, pp. 321-329.

Rényi, A. (1961), *On measures of entropy and inormation*, Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, University of California, Berkley Press, pp. 547-561.

Robbins, H.E. (1968), *Estimating the total probability of the unobserved outcomes of an experiment*, Annals of Mathematical Statistics, 39, pp.256-257.

Rogers, J.A. and Hsu, J.C. (2001), *Multiple Comparisons of Biodiversity*, Biometrical Journal, 43, 5, pp. 617-625.

Serfling, J. Robert (1980), *Approximation Theorems of Mathematical Statistics*

Simpson, E.H. (1949), *Measurement of Diversity*, Nature, Vol. 163, p.688.

Starr, N. (1979), *Linear estimation of probability of discovering a new species*, Annals of Statistics, 7, pp.644-652.

Thisted, R. and Efron, B. (1987), *Did Shakespeare write a newly-discovered poem?*, Biometrika, 74, pp.445-455.

Wang, J.Z. and Lindsay, B.G. (2005), *A penalized nonparametric maximum likelihood approach to species richness estimation*, JASA, Vol. 100, No.471, pp. 942-959.

Zhang, C.-H. (2005), *Estimation of sums of random variables: examples and information bounds*, Annals of Statistics, 33, pp.2022-2041.

Zhang Z.Y. and Huang H.W. (2007), *Turing's formula revisited*, Journal of Quantitative Linguistics, 14, 2-3, pp.222-241.

Zhang Z.Y. and Huang H.W. (2008), *A sufficient normality condition for Turing's formula*, Journal of nonparametric Statistics.

Zhang, C.-H. and Zhang, Z. (2009). *Asymptotic normality of a nonparametric estimator of sample coverage*, Annals of Statistics, (to appear).