
ELSEVIER
Volume 140
Issue 7
July 2010
ISSN 0378-3758
journal of statistical planning and inference
For Contents see page iii
Available online at www.sciencedirect.com
ScienceDirect

Journal of Statistical Planning and Inference 140 (2010) 1731–1738

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Re-parameterization of multinomial distributions and diversity indices

Zhiyi Zhang *, Jun Zhou

Department of Mathematics and Statistics, University of North Carolina at Charlotte, 9201 University City Blvd., Charlotte, NC 28223, USA

A R T I C L E I N F O

Article history:
Received 15 October 2008
Received in revised form
13 March 2009
Accepted 22 December 2009
Available online 4 January 2010

MSC:
primary, 62f10
62F12
62G05
62G20
secondary, 62F15

Keywords:
Generalized Simpson's biodiversity indices
Umvue
Asymptotic normality
Asymptotic efficiency

A B S T R A C T

It is shown in this paper that the parameters of a multinomial distribution may be re-parameterized as a set of generalized Simpson's diversity indices. There are two important elements in the generalization: (1) Simpson's diversity index is extended to populations with infinite species; (2) weighting schemes are incorporated. A class of unbiased estimators for the generalized Simpson's biodiversity indices is proposed. Asymptotic normality is established for the estimators. Both the unbiasedness and the asymptotic normality of the estimators hold for all three cases of the number of species in the population: infinite, finite and known, and finite but unknown. In the case of a population with a finite number of species, known or unknown, it is also established that the proposed estimators are uniformly minimum variance unbiased and are asymptotically efficient.

1. Introduction and summary

Consider a multinomial probability distribution with infinite categories indexed by a positive integer s, i.e., $\{p_s\}=\{p_s; s\geq 1\}$ where p_s may be viewed as the proportion of sth 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

$$\zeta_{1,1}=1-\lambda=\sum_{s=1}^{S} p_s q_s, \tag{1.1}$$

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:

$$\zeta_{u,v}=\sum p_s^u q_s^v, \tag{1.2}$$

* Corresponding author. Tel.: +1 704 687 4549; fax: +1 704 687 6415.
E-mail address: zzhang@uncc.edu (Z. Zhang).

0378-3758/$ - see front matter
doi:10.1016/j.jspi.2009.12.023

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 unless otherwise specified. Eq. (1.2) may be viewed as a weighted version of (1.1), 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, etc.

In the literature of biodiversity, there exists a vast collection of indices. While all are designed to measure species richness of some sort in a population, these indices can loosely 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, 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) and Mao (2007). 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 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 along the line of 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 an adjustment aiming at reducing the bias by various techniques. As a result, the adjusted estimators become more complex in form and their associated 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 correct asymptotic distributional characteristics were derived except in a naive approach (the replicate approach) in which the *iid* sample of size rn is arbitrarily split into r *iid* sub-samples of size n. The asymptotic normality was then achieved by allowing n to increase indefinitely. 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 probability and for all non-uniform $\{p_s\}$ when $\{p_s\}$ contains finitely many species with positive probability. 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. The paper ends with several remarks and miscellaneous results in the last section.

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 $\mathbf{Z}' = M(\mathbf{P}') \subset R^\infty$:

$$M: \{p_s\} \to \{\zeta_u\}. \tag{2.1}$$

Theorem 2.1. *M in* (2.1) *is injective.*

Proof. For every $\{p_s\} \in \mathbf{P}'$, $M(\{p_s\})$ is unique. If 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

$$\sum_{s \geq s_0} p_s^u = \sum_{s \geq s_0} q_s^u \tag{2.2}$$

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

$$\lim_{u \to \infty} \frac{\sum_{s \geq s_0} p_s^u}{p_{s_0}^u} = r_p \geq 1 \quad \text{and} \quad \lim_{u \to \infty} \frac{\sum_{s \geq s_0} q_s^u}{q_{s_0}^u} = r_q \geq 1, \tag{2.3}$$

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 (2.2),

$$\frac{\sum_{s \geq s_0} p_s^u}{p_{s_0}^u} = \frac{\sum_{s \geq s_0} q_s^u}{q_{s_0}^u}\left(\frac{q_{s_0}}{p_{s_0}}\right)^u. \tag{2.4}$$

The right side of (2.4) approaches 0 or ∞ if $p_{s_0} \neq q_{S_0}$, which contradicts (2.3). Therefore s_0 does not exist and $\{p_s\} = \{q_s\}$. $\square$

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 2.1 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 $\zeta_{1,1}$.

3. Estimators

Let X_i, $i = 1, \ldots, n$ 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 sth species found in the sample. The following is the proposed estimator for $\zeta_{u,v}$:

$$Z_{u,v} = \binom{n}{u+v}^{-1}\binom{u+v}{u}^{-1}\sum_{s \geq 1}\left[1_{[Y_s \geq u]}\binom{Y_s}{u}\binom{n-Y_s}{v}\right]. \tag{3.1}$$

$Z_{u,v}$ 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

$$\begin{aligned}
Z_{1,1} &= \frac{n}{n-1}\sum \hat{p}_s(1-\hat{p}_s),\\
Z_{2,0} &= \frac{n}{n-1}\sum 1_{[\hat{p}_s \geq 2/n]}\hat{p}_s(\hat{p}_s-1/n),\\
Z_{3,0} &= \frac{n^2}{(n-1)(n-2)}\sum 1_{[\hat{p}_s \geq 3/n]}\hat{p}_s(\hat{p}_s-1/n)(\hat{p}_s-2/n),\\
Z_{2,1} &= \frac{n^2}{(n-1)(n-2)}\sum 1_{[\hat{p}_s \geq 2/n]}\hat{p}_s(\hat{p}_s-1/n)(1-\hat{p}_s),\\
Z_{1,2} &= \frac{n^2}{(n-1)(n-2)}\sum 1_{[\hat{p}_s \geq 1/n]}\hat{p}_s(1-\hat{p}_s)(1-1/n-\hat{p}_s).
\end{aligned} \tag{3.2}$$

$Z_{u,v}$ 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 $\{X_1, \ldots, X_m\}$, 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 $X_1, \ldots, X_n$ be an *iid* sample under a distribution F, $\theta = \theta(F)$ be a parameter of interest, $h(X_1, \ldots, X_m)$ where $m < n$ be a symmetric kernel satisfying $E_F\{h(X_1, \ldots, X_m)\} = \theta(F)$, $U_n = U(X_1, \ldots, X_n) = \binom{n}{m}^{-1}\sum_k h(X_{i_1}, \ldots, X_{i_m})$ where the summation $\sum_k$ is over all possible sub-samples of size m from the sample of size n, $h_1(x_1) = E_F\{h(x_1, X_2, \ldots, X_m)\}$ 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 3.1. *If $E_F\{h^2\}<\infty$ and $\sigma_1^2>0$, then $\sqrt{n}(U_n-\theta)\xrightarrow{d}\mathrm{N}(0,m^2\sigma_1^2)$.*

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 $h=h(X_1,\ldots,X_m)=(C_m^u)^{-1}N_u$. Let $\mathbf{p}=\{p_s\}$. Suppose $u\geq 1$ and $v\geq 1$. Given $X_1=x_1$,

$$\begin{aligned}
C_m^u h_1(x_1) &= C_m^u E_{\mathbf{P}}\{h(x_1,X_2,\ldots,X_m)\}=E_{\mathbf{P}}\{N_u|X_1=x_1\}\\
&= \sum 1_{[x_{1s}=1]}C_{m-1}^{u-1}p_s^{u-1}q_s^v+\sum 1_{[x_{1s}=0]}C_{m-1}^u p_s^u q_s^{v-1}\\
&= \sum C_{m-1}^u p_s^u q_s^{v-1}+\sum 1_{[x_{1s}=1]}C_{m-1}^u p_s^{u-1}q_s^{v-1}\left(q_s\frac{u}{v}-p_s\right)\\
&= C_{m-1}^u\sum p_s^u q_s^{v-1}+C_{m-1}^u\sum 1_{[x_{1s}=1]}p_s^{u-1}q_s^{v-1}\left(q_s\frac{u}{v}-p_s\right),(C_m^u)^2\sigma_1^2(u,v)\\
&= (C_m^u)^2 Var_{\mathbf{P}}\{h_1(X_1)\}=(C_{m-1}^u)^2 Var_{\mathbf{P}}\left\{\sum 1_{[x_{1s}=1]}p_s^{u-1}q_s^{v-1}\left(q_s\frac{u}{v}-p_s\right)\right\}\\
&= (C_{m-1}^u)^2\left\{E_{\mathbf{P}}\left[\sum 1_{[x_{1s}=1]}p_s^{u-1}q_s^{v-1}\left(q_s\frac{u}{v}-p_s\right)\right]^2-\left[\sum p_s^u q_s^{v-1}\left(q_s\frac{u}{v}-p_s\right)\right]^2\right\}\\
&= (C_{m-1}^u)^2\left\{\sum p_s^{2u-1}q_s^{2v-2}\left(q_s\frac{u}{v}-p_s\right)^2-\left[\sum p_s^u q_s^{v-1}\left(q_s\frac{u}{v}-p_s\right)\right]^2\right\}\\
&= \frac{u^2}{v^2}(C_{m-1}^u)^2\sum p_s^{2u-1}q_s^{2v}-\frac{2u}{v}(C_{m-1}^u)^2\sum p_s^{2u}q_s^{2v-1}+(C_{m-1}^u)^2\sum p_s^{2u+1}q_s^{2v-2}-(C_{m-1}^u)^2\left(\frac{u}{v}\sum p_s^u q_s^v-\sum p_s^{u+1}q_s^{v-1}\right)^2\\
&= \frac{u^2}{v^2}(C_{u+v-1}^u)^2\zeta_{2u-1,2v}-\frac{2u}{v}(C_{u+v-1}^u)^2\zeta_{2u,2v-1}+(C_{u+v-1}^u)^2\zeta_{2u+1,2v-2}-(C_{u+v-1}^u)^2\left(\frac{u}{v}\zeta_{u,v}-\zeta_{u+1,v-1}\right)^2\geq 0. \qquad (3.3)
\end{aligned}$$

The last inequality in (3.3) 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 3.1. 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 3.1 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

$$\sigma_1^2(u,0)=Var_{\mathbf{P}}\{h_1(X_1)\}=\sum p_s^{2u-1}-\left(\sum p_s^u\right)^2=\zeta_{2u-1,0}-\zeta_{u,0}^2\geq 0. \qquad (3.4)$$

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

Thus the following theorem is established.

Theorem 3.1. *If $\{p_s\}$ is a non-uniform multinomial distribution, then for any given pair of positive integers u and v, $Z_{u,v}$ in (3.1), $\zeta_{u,v}$ in (1.2), $\sigma_1^2(u,v)$ in (3.3), and $\sigma_1^2(u,0)$ in (3.4),*

$$\sqrt{n}(Z_{u,v}-\zeta_{u,v})\xrightarrow{d}\mathrm{N}(0,(u+v)^2\sigma_1^2(u,v)) \quad \textit{and} \quad \sqrt{n}(Z_{u,0}-\zeta_{u,0})\xrightarrow{d}\mathrm{N}(0,u^2\sigma_1^2(u,0)). \qquad (3.5)$$

Theorem 3.1 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 (3.3), (3.4) and Theorem 3.1, it is easily seen that when $u\geq 1$ and $v\geq 1$

$$\hat{\sigma}_1^2(u,v)=\left(\frac{v}{u+v}\right)^2\left[\frac{u^2}{v^2}Z_{2u-1,2v}-\frac{2u}{v}Z_{2u,2v-1}+Z_{2u+1,2v-2}-\left(\frac{u}{v}Z_{u,v}-Z_{u+1,v-1}\right)^2\right] \quad \text{and} \quad \hat{\sigma}_1^2(u,0)=Z_{2u-1,0}-Z_{u,0}^2 \qquad (3.6)$$

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 3.1. *If the condition of Theorem 3.1 is satisfied, then for any given pair of positive integers u and v, $Z_{u,v}$ in (3.1), $\zeta_{u,v}$ in (1.2), $\hat{\sigma}_1^2(u,v)$ and $\hat{\sigma}_1^2(u,0)$ in (3.6),*

$$\frac{\sqrt{n}(Z_{u,v}-\zeta_{u,v})}{(u+v)\hat{\sigma}_1(u,v)}\xrightarrow{d}\mathrm{N}(0,1) \quad \textit{and} \quad \frac{\sqrt{n}(Z_{u,0}-\zeta_{u,0})}{u\hat{\sigma}_1(u,0)}\xrightarrow{d}\mathrm{N}(0,1). \qquad (3.7)$$

As a case of special interest when $u=v=1$, the computational formula of $Z_{1,1}$ is given in (3.2) and

$$\frac{\sqrt{n}(Z_{1,1}-\zeta_{1,1})}{2\hat{\sigma}_1(1,1)}\xrightarrow{d}\mathrm{N}(0,1), \qquad (3.8)$$

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 (3.2). Eq. (3.8) 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 deemed 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 (3.1):

$$Z_{u,v}=\frac{n^{u+v}[n-(u+v)]!}{n!}\sum_{s=1}^{S}\left\{1_{[\hat{p}_s\ge u/n]}\prod_{i=0}^{u-1}\left(\hat{p}_s-\frac{i}{n}\right)\left[1_{[v=0]}+1_{[v\ge 1]}\prod_{j=0}^{v-1}\left(1-\hat{p}_s-\frac{j}{n}\right)\right]\right\}. \tag{3.9}$$

Since the coefficient in front of the summation in (3.9) converges to 1 as $n\to\infty$, it is only to show that

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

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

$$\sqrt{n}\left\{\sum_{s=1}^{S}\left\{1_{[\hat{p}_s\ge u/n]}\prod_{i=0}^{u-1}\left(\hat{p}_s-\frac{i}{n}\right)\left[1_{[v=0]}+1_{[v\ge 1]}\prod_{j=0}^{v-1}\left(1-\hat{p}_s-\frac{j}{n}\right)\right]\}-\hat{\zeta}_{u,v}^{(1)}\right\}-\sqrt{n}\hat{\zeta}_{u,v}^{(2)}\xrightarrow{p}0. \tag{3.10}$$

It is to show that each of the two terms in (3.10) converges to zero in probability.

First consider the case of $v=0$. $\prod_{i=0}^{u-1}(\hat{p}_s-i/n)$ 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, $k_2\ge 1$ and $k_3\ge 1$ are finite fixed integers. Since

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

the first term of (3.10) converges to zero in probability. The second terms of (3.10) 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 $u\ge 2$ since there are at most n terms in the sum and

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

Next consider the case of $v\ge 1$. $\prod_{i=0}^{u-1}(\hat{p}_s-i/n)\prod_{j=0}^{v-1}(1-\hat{p}_s-j/n)$ 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, $k_2\ge 1$, $k_3\ge 1$, and $k_4\ge 1$ are finite fixed integers. Since

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

the first term of (3.10) converges to zero in probability. The second term of (3.10) 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 $u\ge 2$ since there are at most n terms in the sum and

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

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

4. Numerical examples

Twelve cases of simulation studies, four distributions by three levels of sample size, are conducted to examine the adequacy of the normal approximation in (3.8). The distributions used in the simulations studies are:

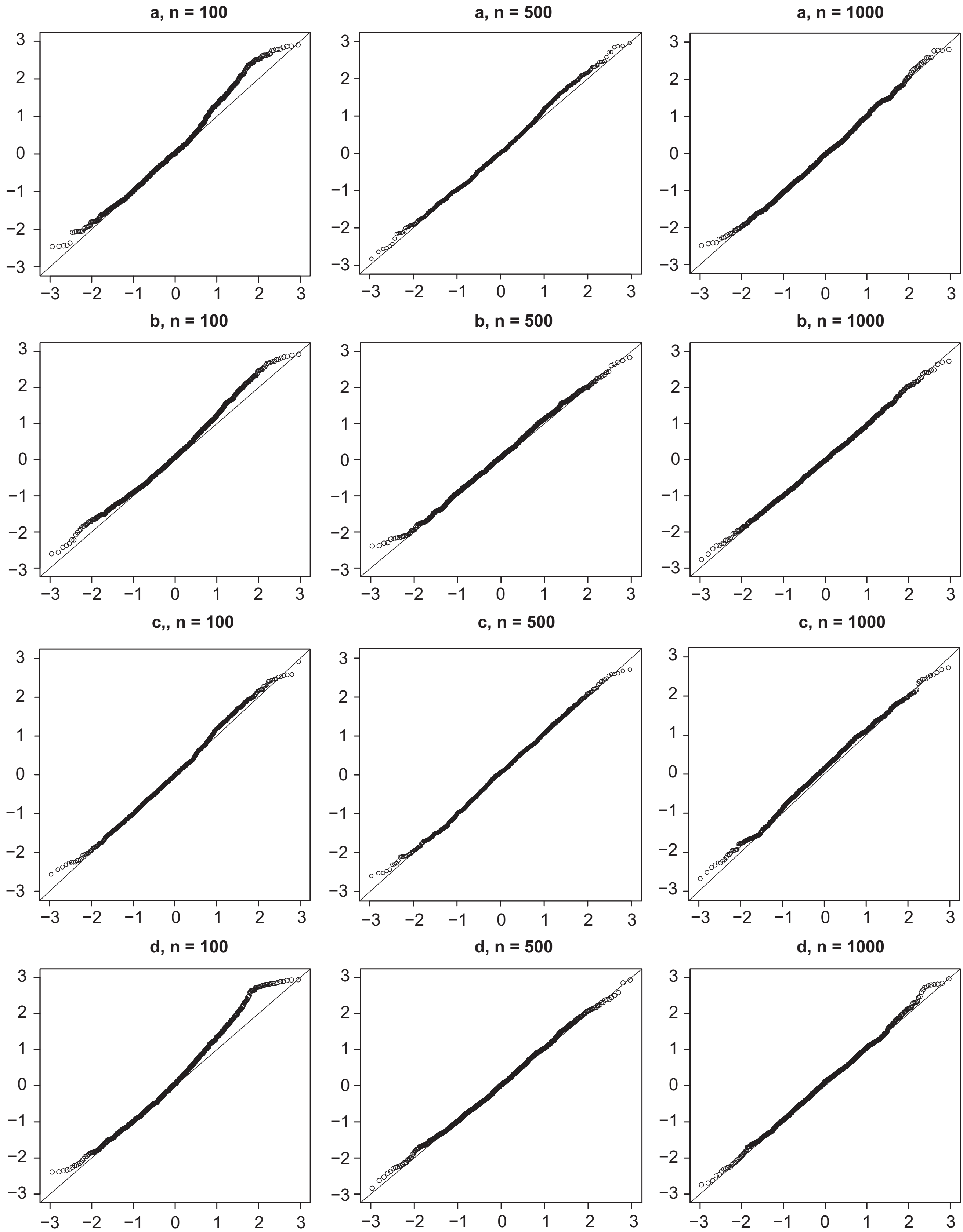

Fig. 1. Q–Q plots for simulated data.

(a) Triangular with $p_s = 0.02(s-0.5)$, $s = 1, \ldots, 10$.
(b) Finite exponential with $p_s = ce^{-s/3}/3$, $s = 1, \ldots, 10$, where $c = (\sum_{s=1}^{10} e^{-s/3}/3)^{-1}$.
(c) Pareto with $p_1 = p_2 = 1/3$, and $p_s = 2/[4(s-1)^2-1]$ for $s \geq 3$.
(d) Exponential with $p_s = e^{-(s-1)/10} - e^{-s/10}$ for $s \geq 1$.

Each distribution is crossed with three levels of sample size, $n = 100$, 500 and 1000. Each simulation study is based on 1000 replications. Q–Q plots against N(0, 1) are given in Fig. 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 (3.8). The range on each axis is from -3 to 3. Columns 1, 2 and 3 in Fig. 1 are corresponding to sample size levels 100, 500 and 1000, respectively.

Fig. 1 indicates that the normality approximation of (3.8) is satisfactory within the range of -3 to 3 when $n = 500$ and 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 (3.8) all seem to have thicker right tails than the standard normal distribution.

5. Miscellaneous

The use of diversity indices is common but is not without skeptics. One usual argument 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 paper offers a panel of estimable indices, which could potentially capture a wider range of diversity.

For (3.1) to be unbiased, $m = u+v$ must be less or equal to the sample size n. However, for (3.7) 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 3.1, and therefore Corollary 3.1, 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 paper.

Definition 5.1. A multi-dimensional parameterization of an underlying distribution, $\{\theta\} \in \Theta$, is said to be sufficient iff $\{\theta\}$ uniquely determines the underlying distribution.

Definition 5.2. 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 5.3. 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 2.1, 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 2.1. 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.

References

Esty, W.W., 1983. A normal limit law for a nonparametric estimator of the coverage of a random sample. The Annal of Statistics 11, 905–912.
Fritsch, K.S., Hsu, J.C., 1999. Multiple comparison of entropies with application to dinosaur biodiversity. Biometrics 55, 1300–1305.
Good, I.J., 1953. The population frequencies of species and the estimation of population parameters. Biometrika 40, 237–264.
Hill, M.O., 1973. Diversity and evenness: a unifying notation and its consequences. Ecology 54, 427–431.
Hoeffding, W., 1948. A class of statistics with asymptotically normal distribution. Annals of Mathematical Statistics 19, 293–325.
Magurran, A.E., 1988. Ecological Diversity and its Measurement. Princeton University Press, Princeton, NJ, USA.
Mao, C.X., 2007. Estimating species accumulation curves and diversity indices. Statistica Sinica 17, 761–774.

Rennolls, K., 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, 321–329.
Rényi, A., 1961. On measures of entropy and information. 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. The Annals of Statistics 39 (1), 256–257.
Rogers, J.A., Hsu, J.C., 2001. Multiple comparisons of biodiversity. Biometrical Journal 43 (5), 617–625.
Simpson, E.H., 1949. Measurement of diversity. Nature 163, 688.
Wang, J.Z., Lindsay, B.G., 2005. A penalized nonparametric maximum likelihood approach to species richness estimation. JASA 100 (471), 942–959.
Zhang, Z., Huang, H., 2007. Turing's formula revisited. Journal of Quantitative Linguistics 14 (2–3), 222–241.
Zhang, Z., Huang, H., 2008. A sufficient normality condition for Turing's formula. Journal of Nonparametric Statistics 20 (5), 431–446.
Zhang, C.-H., Zhang, Z., 2009. Asymptotic normality of a nonparametric estimator of sample coverage. Annals of Statistics 37 (5A), 2582–2595.