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.

© 2009 Elsevier B.V. All rights reserved.

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 $s$th species in a population. Simpson (1949) defined a biodiversity index $\lambda = \sum_{s=1}^{\infty} p_s^2$ for a population with a finite number of species $S$, which has an equivalent form

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

(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:

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

(1.2)
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 unless otherwise specified. Eq. (1.2) may be viewed as a weighted version of (1.1), e.g., \( \zeta_{s,1} \) loads higher weight on minor species (those with smaller \( p_s \)'s), and \( \zeta_{s,2} \) 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 and Zhang (2009) for its statistical properties. In the second category, many different diversity indices have been proposed.

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

\[
M: \{p_i\} \rightarrow \{\zeta_u\}.
\]

(2.1)

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

**Proof.** For every \( \{p_i\} \in P, M(\{p_i\}) \) is unique. If suffices to show that, for every \( \{\zeta_u\} \in Z, M^{-1}(\{\zeta_u\}) \) is unique. Suppose that there existed two sequences, \( \{p_i\} \) and \( \{q_i\} \), in \( P \) satisfying \( \sum p^{u}_{s} = \sum q^{u}_{s} \) for all \( u \geq 1 \). Let \( s_0 = \min(s; p_s \neq q_s) \). If \( s_0 \) does not exist, then \( \{p_i\} = \{q_i\} \). Consider the mapping from \( P \) to \( Z = M(P) \subset R^\infty \):

\[
M: \{p_i\} \rightarrow \{\zeta_u\}.
\]

(2.1)

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

**Proof.** For every \( \{p_i\} \in P, M(\{p_i\}) \) is unique. If suffices to show that, for every \( \{\zeta_u\} \in Z, M^{-1}(\{\zeta_u\}) \) is unique. Suppose that there existed two sequences, \( \{p_i\} \) and \( \{q_i\} \), in \( P \) satisfying \( \sum p^{u}_{s} = \sum q^{u}_{s} \) for all \( u \geq 1 \). Let \( s_0 = \min(s; p_s \neq q_s) \). If \( s_0 \) does not exist, then \( \{p_i\} = \{q_i\} \). Consider the mapping from \( P \) to \( Z = M(P) \subset R^\infty \):

\[
M: \{p_i\} \rightarrow \{\zeta_u\}.
\]

(2.1)

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

**Proof.** For every \( \{p_i\} \in P, M(\{p_i\}) \) is unique. If suffices to show that, for every \( \{\zeta_u\} \in Z, M^{-1}(\{\zeta_u\}) \) is unique. Suppose that there existed two sequences, \( \{p_i\} \) and \( \{q_i\} \), in \( P \) satisfying \( \sum p^{u}_{s} = \sum q^{u}_{s} \) for all \( u \geq 1 \). Let \( s_0 = \min(s; p_s \neq q_s) \). If \( s_0 \) does not exist, then \( \{p_i\} = \{q_i\} \). Consider the mapping from \( P \) to \( Z = M(P) \subset R^\infty \):

\[
M: \{p_i\} \rightarrow \{\zeta_u\}.
\]

(2.1)

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

**Proof.** For every \( \{p_i\} \in P, M(\{p_i\}) \) is unique. If suffices to show that, for every \( \{\zeta_u\} \in Z, M^{-1}(\{\zeta_u\}) \) is unique. Suppose that there existed two sequences, \( \{p_i\} \) and \( \{q_i\} \), in \( P \) satisfying \( \sum p^{u}_{s} = \sum q^{u}_{s} \) for all \( u \geq 1 \). Let \( s_0 = \min(s; p_s \neq q_s) \). If \( s_0 \) does not exist, then \( \{p_i\} = \{q_i\} \). Consider the mapping from \( P \) to \( Z = M(P) \subset R^\infty \):

\[
M: \{p_i\} \rightarrow \{\zeta_u\}.
\]

(2.1)

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

**Proof.** For every \( \{p_i\} \in P, M(\{p_i\}) \) is unique. If suffices to show that, for every \( \{\zeta_u\} \in Z, M^{-1}(\{\zeta_u\}) \) is unique. Suppose that there existed two sequences, \( \{p_i\} \) and \( \{q_i\} \), in \( P \) satisfying \( \sum p^{u}_{s} = \sum q^{u}_{s} \) for all \( u \geq 1 \). Let \( s_0 = \min(s; p_s \neq q_s) \). If \( s_0 \) does not exist, then \( \{p_i\} = \{q_i\} \). Consider the mapping from \( P \) to \( Z = M(P) \subset R^\infty \):

\[
M: \{p_i\} \rightarrow \{\zeta_u\}.
\]

(2.1)

**Theorem 2.1.** \( M \) in (2.1) is injective.
where \( r_p \) and \( r_q \) are multiplicities of \( p_i \)'s with the same value as \( p_{i0} \) and of \( q_i \)'s with the same value as \( q_{i0} \), respectively. But by (2.2),

\[
\frac{\sum_{s \geq s_{0}} p_{i}^p}{p_{i0}^p} = \frac{\sum_{s \geq s_{0}} q_{i}^q}{q_{i0}^q},
\]

(2.4)

The right side of (2.4) approaches 0 or \( \infty \) if \( p_{i0} \neq q_{i0} \), which contradicts (2.3). Therefore \( s_{0} \) does not exist and \( (p_{i}) = (q_{i}) \). \( \square \)

It is to be noted that the monotonicity condition on \( (p_{i}) \) cannot be further relaxed. This is because \( (|v_{s}|) \) is invariant under any permutation of the index set \( (s) \) and \( (p_{i}) \) is not. The one-to-one correspondence between \( P \) and \( Z \) via \( M \) is and can only be established under the monotonicity condition.

Theorem 2.1 has an intriguing implication: the complete knowledge of \( (p_{i}) \) up to a permutation and the complete knowledge of \( (|v_{s}|) \) are equivalent. On the other hand, letting \( Z = (|v_{s}|; u \geq 1, v > 0) \), each member of \( Z \) is a linear combination of finite members of \( Z \). Therefore the complete knowledge of \( (p_{i}) \) up to a permutation and the complete knowledge of \( (|v_{s}|) \) 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 \( (p_{i}) \). \( X_i \) may be written as \( X_i = (X_{is}; s \geq 1) \) where for every \( i \), \( X_{is} \) takes 1 only for one \( s \) and 0 for all other \( s \) values. Let \( Y_i = \sum_{s=1}^{\infty} X_{is} \) and \( \bar{p}_i = Y_i/n \). \( Y_i \) is the number of observations of the \( s \)th species found in the sample. The following is the proposed estimator for \( \zeta_{u,v}^p \):

\[
Z_{u,v}^p = \left( \frac{n}{u+v} \right)^{-1} \left( \frac{u+v}{u} \right)^{-1} \sum_{s=1}^{v} \left[ \frac{1}{u} \sum_{y=1}^{u} \left( \frac{Y_i}{u} \right)^{(n-Y_i)} \right].
\]

(3.1)

\( Z_{u,v}^p \) is a function of \( (Y_i; s \geq 1) \) and hence of \( (\bar{p}_i) = (\bar{p}_s; s \geq 1) \). For a few special pairs of \( u \) and \( v \), \( Z_{u,v}^p \) reduces to

\[
\begin{align*}
Z_{1,1} &= \frac{n}{n-1} \sum \bar{p}_i (1-\bar{p}_i), \\
Z_{2,0} &= \frac{n^2}{(n-1)(n-2)} \sum \left( 1-\bar{p}_i \right), \\
Z_{3,0} &= \frac{n^2}{(n-1)(n-2)} \sum \left( 1-\bar{p}_i \right), \\
Z_{2,1} &= \frac{n^2}{(n-1)(n-2)} \sum \left( 1-\bar{p}_i \right), \\
Z_{1,2} &= \frac{n^2}{(n-1)(n-2)} \sum \left( 1-\bar{p}_i \right).
\end{align*}
\]

(3.2)

\( Z_{u,v}^p \) is an unbiased estimator of \( \zeta_{u,v}^p \). This fact is established by a U-statistic construction of the estimator. Let \( m = u + v \). For every sub-sample of size \( m \), say \( (X_{1s}, \ldots, X_{ms}) \), consider the number of species in the population that are represented exactly \( u \) times in the sub-sample, i.e., \( N_u = \sum_{s=1}^{\infty} \left( \sum_{r=1}^{u} X_{rs} = u \right) \).

\[
E(N_u) = \sum_{s=1}^{m} p \left( \sum_{r=1}^{u} X_{rs} = u \right) = \sum_{s=1}^{m} \left( \frac{m!}{u!} \right) p^u q_s^r.
\]

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

\[
\tilde{Z}_{u,v}^p = \left( \frac{n}{u+v} \right)^{-1} \left( \frac{u+v}{u} \right)^{-1} \sum_{k=1}^{K} N_u^{(k)},
\]

where \( k \) indexes a particular sub-sample is an unbiased estimator of \( \zeta_{u,v}^p \). 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 \( \sum_{k=1}^{K} N_u^{(k)} \) is and can only be established under the monotonicity condition.

The above U-statistic construction paves the path for establishing the asymptotic normality of \( Z_{u,v}^p \). Let \( X_1, \ldots, X_m \) 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 \( E_F(h(X_1, \ldots, X_m)) = \theta(F), U_m = U(X_1, \ldots, X_m) = \left( \frac{m}{n} \right)^{-1} \sum_{k=1}^{K} h(X_1, \ldots, X_m) \) 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 \( \sigma_1^2 = \text{Var}_F(h_1(X_1)) \). The following lemma is by Hoeffding (1948).
Lemma 3.1. If $E_p(h^2) < \infty$ and $\sigma^2 > 0$, then $\sqrt{n}(U_n - \theta) \overset{d}{\rightarrow} N(0, m^2 \sigma^2)$.

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

$$C_m h(x_1) = C_m^{-1} E_p(h(x_1, X_2, \ldots, X_m)) = E_p[N_u | X_1 = x_1]$$

$$= \sum C_m^{u-1} p_{x_1} q_{x_1}^{u-1} + \sum C_m^{u-1} p_{x_1} q_{x_1}^{u-1} (q_{x_1} u_p - p_0)$$

$$= \sum C_m^{u-1} p_{x_1} q_{x_1}^{u-1} + \sum C_m^{u-1} p_{x_1} q_{x_1}^{u-1} (q_{x_1} u_p - p_0)$$

$$= C_m^{-1} Var_p(h(x_1)) = C_m^{-1} \sum C_m^{u-1} p_{x_1} q_{x_1}^{u-1} (q_{x_1} u_p - p_0)$$

$$= (C_m^{-1})^2 \left\{ \sum C_m^{u-1} p_{x_1} q_{x_1}^{u-1} (q_{x_1} u_p - p_0)^2 \right\}$$

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 $p_i$ are equal. Furthermore, since $N_u$ is bounded for every fixed $m$, $E_{|p|}(h^2) < \infty$ is obviously true.

The following definition helps to simplify the subsequent presentation.

Definition 3.1. A multinomial distribution $p_i = (p_i; s \geq 1)$ is said to be uniform if all the non-zero probabilities of $p_i$ are identical.

Definition 3.1 implies that $p_i$ must not be a uniform distribution if it has infinitely many non-zero probabilities.

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

$$\sigma^2(u, 0) = \text{Var}_p(h_1(X_1)) = \sum p_{x_1} q_{x_1} (q_{x_1} u_p - p_0)^2 \geq 0$$

The strict inequality holds for all cases except when $p_i$ is uniform.

Thus the following theorem is established.

Theorem 3.1. If $|p_i|$ 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), $\sigma^2(u, v)$ in (3.3), and $\sigma^2(u, 0)$ in (3.4),

$$\sqrt{n}Z_{u,v} - \zeta_{u,v} \overset{d}{\rightarrow} N(0, \sigma^2(u, v)) \quad \text{and} \quad \sqrt{n}Z_{u,0} - \zeta_{u,0} \overset{d}{\rightarrow} N(0, u^2 \sigma^2(u, 0))$$

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

$$\sigma^2(u, v) = \frac{v}{u + v} \left[ \frac{u^2}{u + v} Z_{2u-1, 2v} - \frac{2v}{u + v} Z_{2u-1, 2v} + Z_{2u+1, 2v-2} - \left( \frac{u^2}{u + v} Z_{u, v} - Z_{u+1, v-1} \right)^2 \right]$$

are consistent estimators of $\sigma^2(u, v)$ and of $\sigma^2(u, 0)$, 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), $\sigma^2(u, v)$ and $\sigma^2(u, 0)$ in (3.6),

$$\frac{\sqrt{n}Z_{u,v} - \zeta_{u,v}}{(u + v) \sigma^2(u, v)} \overset{d}{\rightarrow} N(0, 1) \quad \text{and} \quad \frac{\sqrt{n}Z_{u,0} - \zeta_{u,0}}{u \sigma^2(u, 0)} \overset{d}{\rightarrow} N(0, 1).$$

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

$$\sqrt{n}(Z_{1,1} - \zeta_{1,1}) \overset{d}{\rightarrow} N(0, 1)$$

where $\zeta_{1,1}$ is such that $4 \zeta_{1,1}^2 = Z_{1,2} - 2Z_{1,1} + Z_{1,0} - (Z_{1,1} - Z_{1,0})^2$ and $Z_{1,1}, Z_{1,0}$, and $Z_{0,0}$ 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 $Z_{1,v}$ is unbiased, by the Lehmann–Scheffe Theorem it suffices to show that $(\hat{p}_i)$ is a set of complete and sufficient statistics under $p_i$. When $S$ is finite and known, under the multinomial assumption, $(\hat{p}_i)$ is complete and sufficient. When $S$ is finite but unknown, $(\hat{p}_i)$ is obviously insufficient. The completeness is established.
by the following argument: by the definition of complete statistics, it is to be shown that for any function \( g(\hat{p}_i) \) satisfying \( \mathbb{E}g(\hat{p}_i) = 0 \) for each \((S, \{p_i\})\) implies \( \mathbb{P}(g(\hat{p}_i) = 0) = 1 \) for each \((S, \{p_i\})\). If \( \mathbb{E}g(\hat{p}_i) = 0 \) for each \((S, \{p_i\})\) then for each fixed \(S\), \( \mathbb{E}g(\hat{p}_i) = 0 \) for each \(p_i\) since \(p_i\) is complete for the multinomial distribution, it follows that \( \mathbb{P}(g(\hat{p}_i) = 1) = 1 \) for each \(p_i\). Now \(S\) is arbitrary, thus one actually has \( \mathbb{E}g(\hat{p}_i) = 0 \) for each \((S, \{p_i\})\) implies \( \mathbb{P}(g(\hat{p}_i) = 0) = 1 \) for each \((S, \{p_i\})\).

\(Z_{u,v}\) is asymptotically efficient when \(S\) is finite. This fact is established by recognizing first that \(\hat{p}_i\) is the maximum likelihood estimator (mle) of \(p_i\), second that \(\zeta_{u,v} = \sum \hat{p}_i^u(1-\hat{p}_i)^v\) is the mle of \(\zeta_{u,v}\), and third that \(\sqrt{n}(Z_{u,v} - \zeta_{u,v}) \to 0\) 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):

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

(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

\[
\sqrt{n} \sum_{s=1}^{S} \left\{ s \left[ 1_{[p_i \geq u/n]} \prod_{i=0}^{u-1} \left( \hat{p}_i - \frac{i}{n} \right) \prod_{j=0}^{v-1} \left( 1 - \hat{p}_i - \frac{j}{n} \right) \right] \right\} \to 0,
\]

or letting \(\hat{Z}_{u,v} = \sum_{s=1}^{S} 1_{[p_i \geq u/n]} \hat{p}_i^u(1-\hat{p}_i)^v\) and \(\sum_{s=1}^{S} 1_{[p_i < u/n]} \hat{p}_i^u(1-\hat{p}_i)^v\):

\[
\sqrt{n} \sum_{s=1}^{S} \left\{ s \left[ 1_{[p_i \geq u/n]} \prod_{i=0}^{u-1} \left( \hat{p}_i - \frac{i}{n} \right) \prod_{j=0}^{v-1} \left( 1 - \hat{p}_i - \frac{j}{n} \right) \right] \right\} \to 0.
\]

(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 \(v = 0\). \(\prod_{s=1}^{S} (\hat{p}_i - s/n)\) may be written as a sum of \(\hat{p}_i^u\) and finitely many other terms each of which has the following form:

\[
k_1 \frac{n^u}{n^v} \hat{p}_i^u,
\]

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

\[
0 \leq \sqrt{n} \sum_{s=1}^{S} 1_{[p_i \geq u/n]} \frac{k_1}{n^v} \hat{p}_i^u \leq \sqrt{n} \sum_{s=1}^{S} 1_{[p_i \geq u/n]} \frac{k_1}{n^v} \hat{p}_i < \sqrt{n} \frac{k_1}{n^v} \to 0 \quad \text{as } n \to \infty,
\]

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 \(\zeta_{u,v} = 0\). It also converges to zero in probability when \(u = 2\) since there are at most \(n\) terms in the sum and

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

Next consider the case of \(v \geq 1\). \(\prod_{s=1}^{S} (\hat{p}_i - s/n)\) \(\prod_{s=1}^{S} (1-\hat{p}_i - s/j/n)\) may be written as a sum of \(\hat{p}_i^u(1-\hat{p}_i)^v\) and finitely many other terms each of which has the following form:

\[
k_1 \frac{n^u}{n^v} \hat{p}_i^u(1-\hat{p}_i)^v,
\]

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

\[
0 \leq \sqrt{n} \sum_{s=1}^{S} 1_{[p_i \geq u/n]} \frac{k_1}{n^v} \hat{p}_i^u(1-\hat{p}_i)^v \leq \sqrt{n} \sum_{s=1}^{S} 1_{[p_i \geq u/n]} \frac{k_1}{n^v} \hat{p}_i < \sqrt{n} \frac{k_1}{n^v} \to 0 \quad \text{as } n \to \infty,
\]

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 \(\zeta_{u,v} = 0\). It also converges to zero in probability when \(u = 2\) since there are at most \(n\) terms in the sum and

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

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 \( p_s = c e^{-s/3}, s = 1, \ldots, 10, \) where \( c = (\sum_{s=1}^{10} e^{-s/3})^{-1}. \)

(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.

By Theorem 2.1, under \( P, \) any one \( 0 / C_26 ) \) is minimally sufficient; and (2) there does not exist a proper subset of \( B, \) \( B \subset B, \) such that \( (0) = (0; \beta \in B) \) is sufficient.

Definition 5.3. Two multi-dimensional parameterizations of an underlying distribution, \( (0) \in \Theta \) and \( (0) \in \Omega, \) are said to be equivalent, denoted by \( (0) \equiv (0), \) iff an one-to-one mapping from \( \Theta \) to \( \Omega \) 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 \( (p_s; s \geq 1, s \neq s_2, s \neq s_1, s \neq s_2) \) for any \( s_1 \geq 1 \) and \( s_2 \geq 1 \) is not sufficient. \( (p_s^2; s \geq 1, s \neq s_0) \) for any fixed \( s > 0 \) is also minimally sufficient.

By Theorem 2.1, under \( P, \) \( (0; u \geq 1) \equiv (p_s; s \geq 1). \) Since \( (0; u \geq 1) \subset (0; u \geq 1, v \geq 0), \) \( (0; u \geq 1, v \geq 0) \equiv (p_s; s \geq 1). \) Similarly since \( (0; u \geq 1) \subset (0; u \geq 1, v \geq 0), \) \( (0; u \geq 1, v \geq 0) \equiv (p_s; s \geq 1). \) 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, \( (0; u \geq 1) \) is not minimally sufficient, which implies that \( (0; z \geq 0) \) is not minimally sufficient. The fact that \( (0; u \geq 1) \) is not minimally sufficient can be seen by the fact that any subsequence of \( (0; u) \) 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 \( (0; u) \) does not exist, since a subsequence of any subsequence will uniquely determine the underlying distribution.

References


