Journal of Statistical Distributions and Applications

METHODOLOGY Open Access

A new diversity estimator

Lukun Zheng[1] and Jiancheng Jiang[2*]

*Correspondence:
jjiang1@uncc.edu
[2]Department of Mathematics and Statistics, UNC Charlotte, 9201 University City Blvd, 28223 Charlotte, USA
Full list of author information is available at the end of the article

Abstract
The maximum likelihood estimator (MLE) of Gini-Simpson's diversity index (GS) is widely used but suffers from large bias when the number of species is large or infinite. We propose a new estimator of the GS index and show its unbiasedness. Asymptotic normality of the proposed estimator is established when the number of species in the population is finite and known, finite but unknown, and infinite. Simulations demonstrate advantages of our estimator over the MLE, and a real example for the extinction of dinosaurs endorses the use of our approach. Mathematics Subject Classification (MSC) codes is 60E05, which refers to distributions: general theory.

Keywords: Diversity measure, Gini-Simpson's index, U Statistics

Introduction

Diversity indices are quantitative measures for both richness, the number of categories, and the degree of the evenness of their relative abundances. See Rao (1982), Ludwig and Reynolds (1988), and Patil and Taillie (1979) for further information. It is important to measure the diversity index of a population. For example, in ecology, a decline in diversity over time may indicate a gradual extinction of an ecosystem, while a rapid decline may indicate an extinction due to some sudden impacts. Based on this, scientists argued that the extinction of the dinosaur is due to a large asteroid impact roughly contemporaneous with the end of the Cretaceous. Gini-Simpson's index (GS), together with Shannon's entropy, are the two best known diversity measures. They are widely used in modern sciences such as ecology, demography, anthropology, information theory, and so on. See Hurlbert (1971), Peet (1974), Hunter and Gaston (1988), and Rogers and Hsu (2001).

Consider a population with K species for which p_i denotes the relative abundance of species i ($i = 1, \ldots, K$) such that $\sum_{i=1}^{K} p_i = 1$. Simpson (1949) proposed the index

$$\lambda = \sum_{i=1}^{K} p_i^2 \tag{1}$$

to measure the degree of concentration for the population. Gini-Simpson's index is defined as

$$GS = \sum_{i=1}^{K} p_i(1 - p_i) = 1 - \lambda. \tag{2}$$

There are also many other indices in literature. See Shannon (1948), Good (1953), Renyi (1961), and Hill (1973) among others. In the literature of biodiversity, according to Ricotta

(2005), there are a “jungle" of biological measures of diversity. For a comprehensive discussion on the various relationships among these indices, one may refer to Rennolls and Laumonier (2006) and Mao (2007).

Let $\{X_i\}_{i=1}^n$ be an iid sample from the population $\{p_k; k = 1, \ldots, K\}$, and f_k the observed frequency of the kth category. Let $\hat{p}_k = \frac{f_k}{n}$ and $\hat{P} = \{\hat{p}_k; k = 1, \ldots, K\}$. The most important estimator of GS is the MLE

$$\widehat{GS} = 1 - \sum_{k=1}^{K} \hat{p}_k^2. \tag{3}$$

When K is finite, MLE is asymptotically normal if the underlying distribution is inhomogeneous and is asymptotically distributed as Chi-square if the underlying distribution is homogeneous. Another closed related estimator is given by

$$\frac{n}{n-1}\left[1 - \sum_{k=1}^{K} \hat{p}_k^2\right] = \frac{n}{n-1}\left[1 - \sum_{k=1}\left(\frac{f_k}{n}\right)^2\right]. \tag{4}$$

Bhargava and Uppulurif (1977) showed that it is unbiased and established its asymptotic distribution.

Although the MLE is asymptotically efficient when K is not large relative to the sample size, it does not work well for large K, especially when K is large or infinite. This is easy to understand, since there are only about n/K observations on average for estimating each parameter, and hence the MLE is inefficient when n/K is small. In fact, $\widehat{GS}$ is inconsistent in the case of $K = \infty$ or $K = K_n$ converging to ∞ too fast, and furthermore one cannot use the modern penalized estimation, for example lasso, to estimate p_k, since there is no sparsity structure here. As it will be shown in this paper, MLE also works for the case $K = \infty$ but under some restrictions. Most of the existing methodologies take some adjustment to deal with this problem but result in very complicated forms with less tractable distributional characteristics. Practical techniques include jackknife and bootstrap, see Fritsch and Hsu (1999). Zhang and Zhou (2010) studied a group of estimators for $\zeta_{u,v}$. Due to these problems, little is known about the asymptotic distributional characteristics except in a naive approach. This motivates us to propose a new approach to estimating the GS index. Our new estimator is unbiased, asymptotically normal and efficient for all the cases about the number of species K.

The remainder of the paper is organized as follows. In “A general birthday problem” section, the birthday problem is generalized to cases with unequal probabilities and infinite categories, and the connection between the generalized birthday problem and the GS index is established. In “The estimator” section, based on the relationship between the generalized birthday problem and the GS index, an unbiased estimator of the GS index is proposed and the asymptotic normality is derived under all the three cases with respect to the number of species in the population. In “Asymptotic properties” section, an empirical study about dinosaur extinction data and a simulation study are employed to demonstrate the performance of our estimator.

A general birthday problem

The Birthday problem is an important example in standard textbooks like Feller (1971). The problem is to find the probability that among n students in a class, no two or more students share the same birthday under the assumption that individuals’ birthdays are independent and that for every individual, all 365 days of the year are equally likely as possible birthdays. It has been generalized in many ways under the uniform probability assumption. See Johnson and Kotz (1977) and Fang (1985) among others. Birthday problems with unequal probabilities are also studied over the years. For recent works, see Joag-Dev and Proschan (1992) and Wagner (2002) among others.

Similar to the Bernoulli trial, we define a categorical trial X as a random experiment with K possible outcomes (categories) with probability distribution $P = \{p_k : k = 1, \ldots, K\}$, where K is finite (known or unknown) or infinite. We call it "a success of category i" if the outcome of a categorical trial belongs to category i. Consider an independent sequence of categorical trials $\{X_i; i = 1, 2, \ldots\}$ in which the probability of success of each category keeps the same for each trial. Let H_m be the number of distinct categories shown up in the first m trials. We assume $m \geq 2$ since it is trivial if $m = 1$. Calculating the probability distribution of H_m is generally referred to as birthday problems with unequal probabilities. See the references above. Let Y_k be the number of successes of the kth category in the first m trials and let $I_k = 1_{(Y_k=0)}$ for $k = 1, \ldots, K$ be the indicator function with $I_k = 1$ if the kth category does not appear in the sample . Then

$$H_m = \sum_{k=1}^{K} (1 - I_k)\,. \tag{5}$$

Theorem 1 *For fixed m and finite or infinite value of K, we have*

$$\begin{aligned} E(H_m) &= \binom{m}{1}\sum_{k=1}^{K} p_k - \binom{m}{2}\sum_{k=1}^{K} p_k^2 + \cdots + \binom{m}{m}(-1)^{m+1}\sum_{k=1}^{K} p_k^m, \\ Var(H_m) &= \sum_{k=1}^{K}(1-p_k)^m\left[1-(1-p_k)^m\right]+2\sum_{1\leq i<j\leq K}\left[(1-p_i-p_j)^m - (1-p_i)^m(1-p_j)^m\right]. \end{aligned}$$

The proof of the theorem is given in the Appendix.

Remark 1 *It is easy to see that $Var(H_m)$ is finite for fixed m. In fact, $Var(H_m) < m^2$.*

Now we are ready to establish the connection between the generalized birthday problem and the GS index. For a categorical trial with K categories and probability distribution $P = \{p_k : k = 1, \ldots, K\}$, we have a population of K species with relative abundances $\{p_k : k = 1, \ldots, K\}$. A random sample of size m from this population corresponds to the first m trials in the independent sequence of categorical trials $\{X_i; i = 1, 2, \ldots\}$. As a result, the first m categorical trials can be equivalently viewed as a random sample of size m from the corresponding population, and consequently H_m represents the number of distinct species in a random sample of size m from the corresponding population. The following theorem shows that $GS = 1 - \sum_{k=1}^{K} p_k^2$ is the same as $E(H_2) - 1$.

Theorem 2 *Consider a population with K species and relative abundances $P = \{p_k; k = 1, \ldots, K\}$. Then,*

$$GS = E(H_2) - 1, \tag{6}$$

where H_2 is number of distinct species in a random sample of size 2.

The theorem is a direct result of Theorem 1 taking $m = 2$ and definition of GS (Eq. (2)). The above theorem indicates that GS is an estimable parameter under population P.

The estimator

Let $\{X_i\}_{i=1}^{n}$ be an iid random sample of size n from population P with finite or infinite value of K. For any sub-sample $\{X_{i_1}, \ldots, X_{i_m} : 1 \leq i_1 < \cdots < i_m \leq n\}$ from this sample,

$H_m(X_{i_1}, \ldots, X_{i_m})$ is the number of distinct species in the sub-sample. Therefore $H_m(X_{i_1}, \ldots, X_{i_m})$ is a symmetric function. Define the following U-statistic

$$Z_{n,m} = \binom{n}{m}^{-1} \sum_{c} H_m(X_{i_1}, \ldots, X_{i_m}), \tag{7}$$

where $\sum_c$ denotes the summation over all the $\binom{n}{m}$ combinations of m distinct elements $\{i_1, \ldots, i_m\}$ from $\{1, 2, \ldots, n\}$. Then $Z_{n,m} \to E(H_m)$ almost surely as $n \to \infty$ based on the asymptotic distribution of the U-statistics in DasGupta (2008). This motivates us to estimate GS by

$$\widehat{GS}_1 = Z_{n,2} - 1. \tag{8}$$

It is easy to verify that $\widehat{GS}_1 = Z_{n,2} - 1$ is always an unbiased estimator of GS. In fact, $H_2 - 1$ is an unbiased estimator of GS by Theorem 2, and $Z_{n,2} - 1$ is the average across all combinatorial selections of size 2 from the full set of observations of $H_2 - 1$ applied to each sub-sample.

Asymptotic properties

Asymptotic properties for MLE

Let's firstly prove the asymptotic normality of $\widehat{GS}$ when $K = \infty$. That is, there are infinitely many species in the population. Assume the probability distribution is $P = \{p_i; i = 1, 2, \ldots\}$ with $p_i \geq p_{i+1}$ for all i and $\sum\limits_{i=1}^{\infty} p_i = 1$. And we have the corresponding Gini-Simpson's index $GS = 1 - \sum\limits_{i=1}^{\infty} p_i^2 = 1 - \lambda$. We have the following result.

Theorem 3 *Let $P = \{p_i; i = 1, 2, \ldots\}$ be the probability distribution of a population with infinite species. Assume that there exits a sequence of $\{N_n\}_{n=1}^{\infty}$ such that $np_{N_n+1,+} \to 0$, then we have the following*

$$\frac{\sqrt{n}\left(\widehat{GS} - GS\right)}{\hat{\sigma}} \xrightarrow{p} N(0, 1)$$

where

$$\hat{\sigma}^2 = 4\left[\sum_{i=i}^{N_n} \hat{p}_k^3 - \left(\sum_{i=1}^{N_n} \hat{p}_i^2\right)^2\right]. \tag{9}$$

The proof is given in the Appendix.

The following theorem is implied by Bhargava and Uppulurif (1977) when K is finite, homogeneous or inhomogeneous.

Theorem 4 *If the underlying population distribution is inhomogeneous, then*

$$\frac{\sqrt{n}\left(\widehat{GS}-GS\right)}{\hat{\sigma}} \xrightarrow{p} N(0,1) \tag{10}$$

where

$$\hat{\sigma}^2 = 4\left[\sum_{k=1}^{K}\hat{p}_k^3 - \left(\sum_{k=1}^{K}\hat{p}_k^2\right)^2\right].$$

If the underlying population distribution is homogeneous, we have

$$nK\left(\widehat{GS}-GS\right) \xrightarrow{d} -\chi^2_{K-1}. \tag{11}$$

Asymptotic properties of $\widehat{GS}_1$

The above U-statistic construction paves the way to establish the asymptotic normality of $Z_{n,2}$. For an iid random sample $\{X_i;\ i = 1,\ldots,n\}$ under the distribution P, $\theta = \theta(P)$ is an estimable parameter and $h(X_1,\ldots,X_m)$ is a symmetric kernel satisfying $E_P\{h(X_1,\ldots,X_m)\} = \theta(P)$. Let $U_n = \binom{n}{m}^{-1}\sum\limits_{c} h(X_{i_1},\ldots,X_{i_m})$ where $\sum_c$ is the summation over the $\binom{n}{m}$ combinations of m distinct elements $\{i_1,\ldots,i_m\}$ from $\{1,\ldots,n\}$. Let $h_1(x_1) = E_P\{h(x_1,X_2,\ldots,X_m)\}$ be the conditional expectation of h given $X_1 = x_1$, and $\sigma_1^2 = Var_P\{h_1(X_1)\}$. Then we have the following proposition by Hoeffding (1948).

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

From Remark 1, $E_P\left(h^2\right) = Var\left(H_2(X_1,X_2)\right) + \left(E\left(H_2(X_1,X_2)\right)\right)^2 \le 4 + GS^2 < \infty$. Note that $h_1(x_1) = E_P\left(2 - I_{X_2=x_1}\right) = 2 - p_{x_1}$. It follows that

$$\sigma_1^2 = Var_P\left(h_1(X_1)\right) = Var_P\left(2 - p_{X_1}\right) = Var_P\left(p_{X_1}\right) = \sum_{k=1}^{K} p_k^3 - \left(\sum_{k=1}^{K} p_k^2\right)^2 \ge 0. \tag{12}$$

The equality holds if and only if the probability distribution $\{p_k : k = 1,\ldots,K\}$ is uniform. Of course, if $K = \infty$, the inequality hold strictly since the distribution can never be uniform. Therefore, we have the following theorem.

Theorem 5 *If the distribution* $\{p_k : k = 1,\ldots,K\}$ *is not uniform, then*

$$\sqrt{n}\left(\widehat{GS}_1 - GS\right) \xrightarrow{d} N\left(0, 4\sigma_1^2\right). \tag{13}$$

Remark 2 *Non-uniform distribution includes two cases: non-uniform finite distributions(*$K < \infty$*) and infinite distributions(*$K = \infty$*).*

By (7), (12), and Theorem 1, it is easy to see that

$$\hat{\sigma}_1^2 = Z_{n,3} - Z_{n,2}^2 + Z_{n,2} - 1 \tag{14}$$

is a consistent estimator of σ_1^2. Hence the following corollary is established.

Corollary 1 *Under the conditions of Theorem* 5, *we have*

$$\frac{\sqrt{n}\left(\widehat{GS}_1 - GS\right)}{2\hat{\sigma}_1} \xrightarrow{d} N(0,1). \tag{15}$$

For homogeneous distributions, we have the following result.

Theorem 6 *If the distribution* $\{p_k : k = 1, \ldots, K\}$ *is homogeneous, then*

$$nK\left(\widehat{GS}_1 - GS\right) \overset{d}{\to} \chi^2_{K-1} - K + 1. \tag{16}$$

The proof is given in Appendix. Compared with the MlE estimator, our estimator is reaches the same effect in homogeneous situation.

Examples and simulation studies

Example 1 *(Dinosaur Extinction) The cause of the extinction of dinosaurs at the end of the Cretaceous period remains a mystery. Among all the theories, it is now widely accepted that it is due to a large asteroid impact at the end of the cretaceous. Sheehan et al. (1991) argued that diversity remained relatively constant throughout the Cretaceous period. The scientists reason that if the disappearance of the dinosaurs was gradual, one should observe a decline in diversity prior to extinction.*

The data were organized by dividing the formation into three equally spaced stratigraphic levels, each of which represented a period of approximately 730,000 years. Fossils were cross-tabulated according to the stratigraphic level and the family to which the dinosaur belonged. Families represented are Cerotopsidae, Hadrosauridae, Hypsilophodontidae, Pachycephalosauridae, Tryrannosauridae, Ornithomimidae, Saurornithoididae, Dromaeosauridae. The summarized data is shown in Table 1 *available in Rogers and Hsu (*2001*).*

Let's denote the true value of GS indices at the Lower, Middle, and Upper level by GS_L, GS_M, *and* GS_U, *respectively. It is interesting to ask if the dinosaur diversity changed.*

To address the questions, we would like to present 95% simultaneous confidence intervals for all the pairwise contrasts: $GS_L - GS_M$, $GS_L - GS_U$, *and* $GS_M - GS_U$.

Using expressions for $\widehat{GS}_1$ *and* $\hat{\sigma}_1^2$ *from the previous section and the normal approximation in our theorems, we obtain simultaneous confidence intervals for all pairwise contrasts. The results are provided in Table* 2

Since all the confidence intervals contain zero, we may infer that all three communities were practically equivalent with respect to the GS index. That is, there is no significant change or decline of the diversity over time. Therefore, our study supports the theory of a sudden extinction of dinosaurs.

Our proposed estimator has advantages over the MLE when the sample size n is not large relative to the number of species K, especially when $K = \infty$. In the following we conduct a simulation study for $K = \infty$. We omit simulations for other scenarios for saving space.

Table 1 Dinosaur counts by family and stratigraphic level

Interval	Counts
Upper	(50, 29, 3, 0, 3, 4, 1, 0)
Middle	(53, 51, 2, 0, 3, 8, 6, 0)
Lower	(19, 7, 1, 0, 2, 0, 3, 0)

Table 2 95% simultaneous confidence intervals for all pairwise contrasts

Contrast	Estimate	Std.Error	Critical value $z_{\alpha/6}$	Lower bound	Upper bound
$GS_L - GS_M$	-0.0474	0.0733	2.3941	-0.2229	0.1281
$GS_L - GS_U$	-0.0525	0.0774	2.3941	-0.2378	0.1328
$GS_M - GS_U$	-0.0050	0.0414	2.3941	-0.1041	0.0941

Example 2 *($K = \infty$) Consider the population $\{p_k = e^{-(k-1)/10} - e^{-k/10} : k \geq 1\}$. It is easy to calculate the true value of GS for this distribution:*

$$GS = 1 - \sum_{k=1}^{\infty} p_k^2 = 0.95004.$$

We generate random samples of size n=10, 50, and 100, and calculate the MLE $\widehat{GS}$ and our proposed estimator $\widehat{GS}_1$, together with their standard deviations(Eqs. (9') *and* (14)*). The simulation is based on 500 replications and the results are obtained by averaging the corresponding estimates in each replication . Also, since it is known that the population distribution is not uniform, we will just apply $\widehat{GS}_1$ due to the reason mentioned before. The simulation results are summarized in Table* 3*.*

From Table 3*, we see that the deviations of the MLEs from the true value $GS = 0.95004$ are much greater than those of our proposed estimates. This is due to the facts that $\widehat{GS}$ has a large bias and that the sample coverage is limited when the sample size is relatively small compared with the number of species. Our proposed estimator, instead, overcome such obstacles since it is an unbiased estimator of GS.And it is also shown that our proposed estimator has smaller variance.*

Discussion

Birthday problem has been studied and extended in different forms and in many different areas. The same is true for diversity measures. The connection between these two topics is established in this paper through H_2 and the mostly used Gini-Simpson's index. There are many other correlated diversity indices in the literature, like Shannon's entropy, Renyi's index. For these indices, we can also find corresponding estimators in a similar way through the result in Theorem 1. The advantage of our approach over the MLE is obvious when the sample size is not large relative to the number of species. There are many other open problems built on this connection between birthday problem and diversity measures. For example, further investigation is needed to study the estimation of mutual information in view of generalized birthday problem. Our approach provides a framework for solving various problems inherited from the diversity measures.

Appendix 1: Proof of Theorem 1

Theorem 1 *For fixed m and finite or infinite value of K, we have*

$$E(H_m) = \binom{m}{1} \sum_{k=1}^{K} p_k - \binom{m}{2} \sum_{k=1}^{K} p_k^2 + \cdots + \binom{m}{m} (-1)^{m+1} \sum_{k=1}^{K} p_k^m,$$

Table 3 Estimates of *GS* for Example 2

$K = \infty$	$n = 10$	$n = 50$	$n = 100$
$\widehat{GS}$	0.8532 (std: 0.0354)	0.9317 (std: 0.0107)	0.9407 (std: 0.0068)
$\widehat{GS}_1$	0.9480 (std: 0.0075)	0.9507 (std: 0.0061)	0.9502 (std: 0.0053)

$$Var(H_m) = \sum_{k=1}^{K}(1-p_k)^m\left[1-(1-p_k)^m\right]+2\sum_{1\le i<j\le K}\left[(1-p_i-p_j)^m-(1-p_i)^m(1-p_j)^m\right].$$

Proof Let's consider the following lemma first. □

Lemma 1 *For the class of random variables* $\{I_k; k=1,\dots,K\}$*, we have*

$$E(I_k) = (1-p_k)^m; \tag{17}$$

$$Var(I_k) = (1-p_k)^m - (1-p_k)^{2m}, \tag{18}$$

$$Cov(I_i, I_j) = (1-p_j-p_k)^m - (1-p_j)^m(1-p_k)^m, \mathit{for}\ i \neq j \tag{19}$$

Lemma 1 can be verified easily.

When K is finite, the following equations are easily established.

$$\begin{aligned}
E(H_m) &= \sum_{k=1}^{K}(1-EI_k)\\
&= \sum_{k=1}^{K}\left(1-(1-p_k)^m\right)\\
&= \sum_{k=1}^{K}\left(\binom{m}{1}p_k - \binom{m}{2}p_k^2 + \cdots + \binom{m}{m}(-1)^{m+1}p_k^m\right)\\
&= \binom{m}{1}\sum_{k=1}^{K}p_k - \binom{m}{2}\sum_{k=1}^{K}p_k^2 + \cdots + \binom{m}{m}(-1)^{m+1}\sum_{k=1}^{K}p_k^m, \text{and}\\
Var\,(H_m) &= Var\left[\sum_{k=1}^{K}(1-I_k)\right]\\
&= \sum_{k=1}^{K}Var\,(1-I_k) + 2\sum_{1\le i<j\le K}Cov(1-I_i, 1-I_j)\\
&= \sum_{k=1}^{K}Var\,(I_k) + 2\sum_{1\le i<j\le K}Cov(I_i, I_j)\\
&= \sum_{k=1}^{K}(1-p_k)^m\left[1-(1-p_k)^m\right]+2\sum_{1\le i<j\le K}\left[(1-p_i-p_j)^m-(1-p_i)^m(1-p_j)^m\right]
\end{aligned}$$

When K is infinite, the above equations are guaranteed by dominated convergence theorem. In fact, we have $H_m \le m$ and $H_m^2 \le m^2$.

Appendix 2: Proof of Theorem 3

Theorem 3 *Let* $P = \{p_i; i = 1, 2, \dots\}$ *be the probability distribution of a population with infinite species. Assume that there exits a sequence of* $\{N_n\}_{n=1}^{\infty}$ *such that* $np_{N_n+1,+} \to 0$*, then we have the following*

$$\frac{\sqrt{n}\left(\widehat{GS}-GS\right)}{\hat{\sigma}} \xrightarrow{p} N(0,1)$$

where

$$\hat{\sigma}^2 = 4\left[\sum_{i=i}^{N_n} \hat{p}_k^3 - \left(\sum_{i=1}^{N_n} \hat{p}_i^2\right)^2\right]. \tag{9'}$$

Proof Now let's consider a sequence of populations with probability distributions $P_N = \{p_1, p_2, \ldots, p_{N-1}, P_{N,+}\}$, where $p_{N,+} = \sum_{i=N}^{\infty} p_i$. The corresponding Gini-Simpson's index is

$$GS_N = 1 - \left(\sum_{i=1}^{N-1} p_i^2 + p_{N,+}^2\right) = 1 - \lambda_N.$$

It is easy to check that

$$\lambda_N \to \lambda$$

as $N \to \infty$.

Let $\{X_i\}_{i=1}^n$ be an iid sample from the population P. The MLE of GS is

$$\widehat{GS} = 1 - \sum_{i=1}^{\infty} \hat{p}_k^2.$$

For fixed N, let's re-label the same sample $\{X_i\}_{i=1}^n$ to another sample $\{Y_i\}_{i=1}^n$ as follows:

$$Y_i = X_i \text{ if } X_i < N$$
$$Y_i = N \text{ if } X_i \geq N$$

Then $\{Y_i\}_{i=1}^n$ can be regarded as a iid sample from P_N with Gini-Simpson's index GS_N. The MLE of GS_N is

$$\widehat{GS}_N = 1 - \left(\sum_{i=1}^{N-1} \hat{p}_i^2 + \hat{p}_{N,+}^2\right).$$

It is easy to see that

$$\widehat{GS}_N - \widehat{GS} \to 0$$

as $N \to \infty$. In fact,

$$\widehat{GS} - \widehat{GS}_N = 0 \tag{20}$$

if $X_i \leq N$ for all $i = 1, 2, \ldots, n$.

Therefore,

$$\begin{aligned} &\sqrt{n}\left(\widehat{GS} - GS\right) \\ &= \sqrt{n}\left(\widehat{GS} - \widehat{GS}_N + \widehat{GS}_N - \lambda_N + \lambda_N - \lambda\right) \\ &= \sqrt{n}\left(\widehat{GS} - \widehat{GS}_N\right) + \sqrt{n}\left(\widehat{GS}_N - \lambda_N\right) + \sqrt{n}(\lambda_N - \lambda) \end{aligned}$$

For any positive integer n, consider a corresponding integer N_n. The probability that all the observations in the sample $\{X_i\}_{i=1}^n$ is less or equal to N_n is

$$\left(1 - p_{N_n+1,+}\right)^n = \left(1 - \sum_{i=N_n+1}^{\infty} p_i\right)^n.$$

Therefore, if

$$\left(1-\sum_{i=N_n+1}^{\infty} p_i\right)^n = \left(1-p_{N_n+1,+}\right)^{\frac{np_{N_n+1,+}}{p_{N_n+1,+}}} = e^{-np_{N_n+1}} \to 1$$

that is,

$$np_{N_n+1,+} \to 0 \tag{21}$$

then all the observations in the sample $\{X_i\}_{i=1}^n$ falls into the first N_n species with probability going to 1 as n increases. In turn,

$$\widehat{GS} - \widehat{GS}_N$$

equal to zero with probability going to 1 due to Eq. (20). Therefore,

$$\sqrt{n}\left(\widehat{GS} - \widehat{GS}_N\right)$$

converge to o with probability going to 1 as n increases.

In addition,

$$\begin{aligned}\sqrt{n}\left(\lambda_{N_n} - \lambda\right) &= \sqrt{n}\left(\sum_{i=1}^{N_n-1} p_i^2 + p_{N_n,+}^2 - \sum_{i=1}^{\infty} p_i^2\right)\\ &= \sqrt{n}\left[\left(\sum_{i=N_n}^{\infty} p_i\right)^2 - \sum_{i=N_n}^{\infty} p_i^2\right]\\ &\leq \sqrt{n}\sum_{i=N_n}^{\infty} p_i^2\\ &\leq \sqrt{n}p_{N_n,+}\end{aligned}$$

Therefore, if $\sqrt{n}p_{N_n,+} \to 0$ which is a weaker condition than (21), we have

$$\sqrt{n}\left(\lambda_{N_n} - \lambda\right) \to 0.$$

Therefore, by Slutsky's theorem, the theorem is proved. □

Appendix 3: Proof of Theorem 6

Theorem 6 *If the distribution* $\{p_k : k = 1, \ldots, K\}$ *is homogeneous, then*

$$nK\left(\widehat{GS}_1 - GS\right) \xrightarrow{d} \chi^2_{K-1} - K + 1. \tag{16'}$$

Proof For an iid random sample $\{X_i;\ i = 1, \ldots, n\}$ under the distribution P, $\theta = \theta(P)$ is an estimable parameter and $h(X_1, \ldots, X_m)$ is a symmetric kernel satisfying $E_P\{h(X_1, \ldots, X_m)\} = \theta(P)$. Let $U_n = \binom{n}{m}^{-1} \sum\limits_{c} h(X_{i_1}, \ldots, X_{i_m})$ where $\sum_c$ is the summation over the $\binom{n}{m}$ combinations of m distinct elements $\{i_1, \ldots, i_m\}$ from $\{1, \ldots, n\}$. Let $h_1(x_1) = E_P\{h(x_1, X_2, \ldots, X_m)\}$ be the conditional expectation of h given $X_1 = x_1$, and $\zeta_1 = Var_P\{h_1(X_1)\}$. Also let $h_2(x_1, x_2) = E_P\{h(x_1, x_2, X_3 \ldots, X_m)\}$ be the conditional expectation of h given $X_1 = x_1, X_2 = x_2$, and $\zeta_2 = Var_P\{h_2(X_1, X_2)\}$. Define

$$\tilde{h}_2 = h_2 - \theta.$$

Then we have the following lemmas by Hoeffding (1948). □

Lemma 2 *If* $E_P(h^2) < \infty$ *and* $\zeta_1 > 0$, *then* $\sqrt{n}\,(U_n - \theta) \xrightarrow{d} N\left(0, m^2\zeta_1\right)$.

Lemma 3 *If* $E_P(h^2) < \infty$ *and* $\zeta_1 = 0 < \zeta_2$, *then* $n\,(U_n - \theta) \xrightarrow{d} \frac{m(m-1)}{2}Y$, *where* Y *is a random variable of the form*

$$Y = \sum_j \lambda_j \left(\chi^2_{1j} - 1\right),$$

where $\chi_{11}, \chi_{12}, \ldots$ *are independent* χ^2_1 *variates and* λ_j*s are the eigenvalues of the following operator on the function space* $L_2(R, P)$*:*

$$Ag(x) = \int_{-\infty}^{\infty} \tilde{h}_2(x,y)g(y)dP(y),\ x \in R, g \in L_2. \tag{22}$$

For our case, we have $\theta = GS + 1$ and the kernal function given as

$$h(x_1, x_2) = V(x_1, x_2) = 2 - I_{x_1 = x_2}. \tag{23}$$

That is,

$$GS + 1 = E_P\{h(X_1, X_2)\}$$

for given population distribution P.

Under the assumption of homogeneous population distribution, $\zeta_1 = 0$. Since

$$h(X_1, X_2) = 2 - I_{X_1 = X_2} = \begin{cases} 1 & \text{if } X_1 = X_2 \\ 2 & \text{if } X_1 \neq X_2 \end{cases}$$

We have

$$\begin{aligned}
\zeta_2 &= Var\,(h(X_1, X_2)) \\
&= E\left(h^2(X_1, X_2)\right) - (E(h(X_1, X_2))^2 \\
&= 1 \cdot P(X_1 = X_2) + 4P(X_1 \neq X_2) - (P(X_1 = X_2) + 2P(X_1 \neq X_2))^2 \\
&= P(X_1 = X_2) + 4P(X_1 \neq X_2) - P^2(X_1 = X_2) - 4P(X_1 = X_2)P(X_1 \neq X_2) - 4P^2(X_1 \neq X_2) \\
&= P(X_1 = X_2)(1 - P(X_1 = X_2)) + 4P(X_1 \neq X_2)\left[1 - P(X_1 = X_2) - P(X_1 \neq X_2)\right] \\
&= P(X_1 = X_2)P(X_1 \neq X_2) \\
&= \sum_{i=1}^{K} p_i^2 \left(1 - \sum_{i=1}^{K} p_i^2\right) > 0
\end{aligned}$$

Also

$$\theta = GS + 1 = 2 - \sum_{i=1}^{K} \frac{1}{K^2} = 2 - \frac{1}{K}.$$

Now let's find the eigenvalues of operator A under the homogeneous distribution. We have $\tilde{h}_2(x,y) = 2 - I_{x=y} - \theta = \frac{1}{K} - I_{x=y}$. And

$$\begin{aligned}
Ag(x) &= \int_{-\infty}^{\infty} \tilde{h}_2 g(y) dP(y) \\
&= \int_{-\infty}^{\infty} \left(\frac{1}{K} - I_{x=y}\right) g(y) dP(y) \\
&= \frac{1}{K^2}\sum_{i=1}^{K} g(i) - \frac{1}{K} g(x) \\
&= \frac{1}{K^2}\sum_{i\neq x} g(i) + \left(\frac{1}{K^2} - \frac{1}{K}\right) g(x)
\end{aligned}$$

Since $g : \{1, 2, \ldots, K\} \to R$, it can be viewed as a vector from R^K. And A is a linear operator on R^K. And the matrix representation of A is

$$A\vec{g} = T\vec{g}$$

where T is a $K \times K$ matrix with $T(i,i) = \frac{1}{K^2} - \frac{1}{K}$ and $T(i,j) = \frac{1}{K^2}$ for $i \neq j$. The matrix T has two eigenvalues $\lambda = 0$ with multiplicity one and $\lambda = -\frac{1}{K}$ with multiplicity $K - 1$.

Therefore due to Lemma 3 and properties of independent Chi-square distributions, theorem is proved

Appendix 4: About the variances of $\widehat{GS}$ and $\widehat{GS}_1$

From section of Asymptotic behaviour for homogeneous case, we get that

$$\zeta_1 = \sum_{i=1}^{K} p_i^3 - \left(\sum_{i=1}^{K} p_i^2\right)^2$$

and

$$\zeta_2 = \sum_{i=1}^{K} p_i^2 \left(1 - \sum_{i=1}^{K} p_i^2\right).$$

By the following lemma by Hoeffding (1948):

Lemma 4 *The variance of U_n is given by*

$$Var_F(U_n) = \binom{n}{m}^{-1} \sum_{c=1}^{m} \binom{m}{c}\binom{n-m}{m-c}\zeta_c \tag{24}$$

Therefore,

$$\begin{aligned}
Var\left(\widehat{GS}_1\right) &= \binom{n}{2}^{-1} (2(n-2)\zeta_1 + \zeta_2) \\
&= \frac{2}{n(n-1)}\left[2(n-2)\left(\sum_{i=1}^{K} p_i^3 - \left(\sum_{i=1}^{K} p_i^2\right)^2\right) + \sum_{i=1}^{K} p_i^2 - \left(\sum_{i=1}^{K} p_i^2\right)^2\right] \\
&= \frac{2}{n(n-1)}\left[2(n-2)\sum_{i=1}^{K} p_i^3 - (2n-3)\left(\sum_{i=1}^{K} p_i^2\right)^2 + \sum_{i=1}^{K} p_i^2\right]
\end{aligned}$$

From Bhargava and Uppuluri (1977), we have

$$Var\left(\widehat{GS}\right) = \frac{(n-1)^2}{n^2}\frac{2}{n(n-1)}\left[2(n-2)\sum_{i=1}^{K}p_i^3 - (2n-3)\left(\sum_{i=1}^{K}p_i^2\right)^2 + \sum_{i=1}^{K}p_i^2\right]$$

Therefore, we have the following theorem.

Theorem 7 *When K is finite, we have*

$$Var\left(\widehat{GS}\right) = \frac{(n-1)^2}{n^2}Var\left(\widehat{GS}_1\right).$$

Authors' contributions

LZ contributed in the following ways: Conception and design of study. Data collection, data analysis and interpretation. Drafting the article. JJ contributed to the paper in the following ways: Critical revision of the article. Final approval of the version to be published. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Department of Mathematics, Tennessee Technological University, 1 William L Jones Dr, TN 38505 Cookeville, USA.
[2]Department of Mathematics and Statistics, UNC Charlotte, 9201 University City Blvd, 28223 Charlotte, USA.

Received: 7 January 2017 Accepted: 6 July 2017
Published online: 15 September 2017

References

Bhargava, N, Uppuluri, VRX: Sampling distributions of Gini's index of diversity. Appl. Math. Campn. **3**, 1–24 (1977)
DasGupta, A: Asymptotic Theory of Statistics and Probability. Springer, New York (2008)
Fang, KT: Occupancy problems. Encyclopedia of Statistical Sciences (Kotz, S, Johnson, NL, eds.) Wiley, New York (1985)
Feller, W: An Introduction to Probability Theory and Its Applications, vol. 1. 2nd ed. Wiley, New York (1971)
Fritsch, KS, Hsu, JC: Multiple comparison of entropies with applications to dinosaur biodiversity. Biometrics. **55**, 1300–1305 (1999)
Good, IJ: The population frequencies of species and the estimation of population parameters. Biometrika. **40**, 237–264 (1953)
Hill, MO: Diversity and evenness: A unifying notation and its consequences. Ecology. **54**, 427–432 (1973)
Hoeffding, W: A class of statistics with asymptotically normal distribution. Ann. Math. Stat. **19**, 293–325 (1948)
Hunter, PR, Gaston, MA: Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. J. Clin. Microbiol. **26**(11), 246–466 (1988)
Hurlbert, SH: The nonconcept of species diversity: A critique and alternative parameters. Ecology. **52**, 577–586 (1971)
Joag-Dev, K, Proschan, F: Birthday problem with unlike probabilities. Am. Mat. Mon. **99**, 10–12 (1992)
Johnson, NL, Kotz, S: Urn models and their application. Wiley, New York (1977)
Ludwig, J, Reynolds, JF: Patterns of the abundance of species: a comparison of two hierarchical models. OIKOS. **53**, 235–241 (1988)
Mao, CX: Estimating species accumulation curves and diversity indices. Stat. Sinica. **17**, 761–774 (2007)
Patil, GP, Taillie, C: An overview of diversity. In: Grassle, JF, Tatil, GP, Smith, WK, Taillie, C (eds.) Ecological Diversity in Theory and Practice, pp. 3–27. International Co-operative Publishing House, Fairland, (1979)
Peet, RK: The measurement of species diversity. Ann. Rev. Ecol. System. **5**, 285–307 (1974)
Rao, CR: Diversity: Its measurement, decomposition, apportionment and analysis. Sankhya. **A44**, 1–22 (1982)
Renyi, A: On measures of entropy and information. In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1. Contributions to the Theory of Statistics. The Regents of the University of California pp. 547–561, (1961)
Rennolls, K, Laumonier, Y: A New Local Estimator of Regional Species Diversity, in Terms of 'Shadow Species', with a Case Study from Sumatra. J. Trop. Ecol. **22**, 321–329 (2006)
Ricotta, C: Through the jungle of biological diversity. Acta. Biotheor. **53**(1), 29–38 (2005)
Rogers, J, Hsu, J: Multiple comparisons of biodiversity. Biometrical. J. **43**, 617–625 (2001)
Shannon, CE: A mathematical theory of communication. Bell Syst. Tech. J. **27**, 379–423 (1948)
Sheehan, PM, et al.: Sudden extinction of the Dinosaurs: Latest Cretaceous, Upper Great Plains, U. S. A. Science 254.5033, 835–839 (1991)
Simpson, EH: Measurement of diversity. Nature. **163**, 688 (1949)
Wagner, D: A generalized birthday problem. In Crypto, vol. 2442, pp. 288-303. Springer-Verlag (2002)
Zhang, Z, Zhou, J: Re-parameterization of multinomial distributions and diversity indices. J. Stat. Plan. Infer. **140**, 1731–1738 (2010)