CHAO, JACKKNIFE AND BOOTSTRAP ESTIMATORS OF SPECIES RICHNESS

CHAVAN KR. SARMAH

ABSTRACT: The species richness is one of the oldest, simplest and instant concepts of characterizing biocommunity diversity in Ecological study. Practically, species richness is the basic statistics used to estimate the true (actual) number of species of a biocommunity (animal or plant) in an area under consideration. In this study, in the light of available literature, we focus on, how to estimate the species richness based on the observed species from the sampled individuals using nonparametric estimation techniques. More specifically, it concentrates on the basic ideas and applications of three prominent nonparametric estimators, namely, Chao, Jackknife and Bootstrap estimate of species richness. This paper has started with the basic concept of these estimators with special emphasis on Bootstrap method and concluded with an exposition of respective estimates of variances.

Keywords: Biocommunity, species richness, nonparametric, Chao, Jackknife and Bootstrap etc.

Mathematics Subject Classification: 62Gxx

1. INTRODUCTION

The species richness is one of the oldest, simplest and instant concepts of characterizing biocommunity diversity in Ecological study [1]. Practically, species richness is the basic statistics used to estimate the true number of species of a biocommunity (animal or plant community) in an area under consideration. The study of species richness of various biocommunities is important for conservation, ecological biodiversity management and environmental policy makings. In most of the Biological and Ecological studies species richness constitutes an unknown parameter of the study community. This unknown parameter must be evaluated by means of sampled individuals because completion of complete enumeration of species often requires extraordinary efforts and is almost unattainable goal.

Theoretically, the total species richness can be determined for any biocommunity because number of species is limited. But in practice, species counting of a particular biocommunity for example, spiders faces the number of operational disadvantages. For such disadvantages of counting total number of species, Ecologist employed the Statistical estimators to estimate species richness of a community in an area. Here
among the statistical estimators only few of the nonparametric estimators namely Chao, Jackknife and Bootstrap of species richness has been discussed and illustrated with purposive data considered in this purpose. It is to be noted that observed count species from the study field (sample) is the lower bound for the true number of species (estimated number of species). The nonparametric estimators included in this study will be broadly discussed immediately after the notation section.

2. Notations Used

$S^*$ is the true number of estimated species for the study area.

$S_{obs}$ is the observed number of species in the sample collected from the study area.

‘$a$’ is the singletons species in the sample i.e. number of species that are represented by only a single individual in the sample.

‘$b$’ is the doubletons species in the sample i.e. number of species that are represented by only two individuals in the sample.

‘$n$’ is the number of sample units (plots of the study area for collection of data).

‘$L$’ is the number of species which appears in only one sample unit. Here the notations are considered with the help of [2].

3. Chao Estimator of Species Richness

Using the concept of rare species Chao [3] derived the simple estimator to carry the most information about the number of unsampled species of an area. Based on the observed number of species and the concept of singletons and doubletons species he proposed a lower bound of species richness as

$$S_{cho}^* = S_{obs} + \frac{a^2}{2b}$$

The Chao estimator $S_{cho}^*$ of species richness breaks down when $b = 0$ i.e. if there are no doubletons species in the sample the above stated Chao estimator will not holds. However a modified version was given later as

$$S_{cho}^* = S_{obs} + \frac{a(a - 1)}{2(b + 1)}$$

This is always obtainable. And the variance of the Chao estimator is given by

$$Var(S_{cho}^*) = b \left[ \left( \frac{R^4}{4} \right) + R^3 + \left( \frac{R}{2} \right)^2 \right]$$

where, $R$ is the ratio of number of singletons to doubletons species.
The standard error of the Chao estimator may be considered as

\[ SE = \sqrt{Var(S_{\text{cho}}^*)} = b \left[ \left( \frac{R}{4} \right)^4 + R^3 + \left( \frac{R}{2} \right)^2 \right] \]

### 4. Jackknife Estimator of Species Richness

The Jackknife method was first introduced by Quenouille [4] and further developed by Tukey [5] as a general nonparametric estimator to reduce the bias of a biased estimator. It has a universal application and one of the most applied fields is Ecology. In this study the Jackknife estimator is applied to estimate the species richness in Ecological or Biological study in which observed species is considered as the bias estimator. The Jackknife is a useful estimator of species richness because it is based on the presence or absence of a species in a given area rather than on the abundance of the species. Heltshe and Forrester [6] and Smith and van Belle [7] were first used the Jackknife estimator to estimate species richness.

To use the Jackknife estimator for species richness data must be sampled at different locations (plots or sampling units say, \( n \)) in the area under consideration for which species richness to be estimated. The procedure of Jackknife estimates of species richness is described in the following steps.

**Step 1:** Delete one of the observations (sample unit or plot) say, ‘\( i \)’, here \( i = 1, 2, \ldots, n \) sample units.

**Step 2:** Compute species richness \( \hat{S}_{-i} \), from all observations (sample units) excluding ‘\( i \)’.

**Step 3:** Calculate pseudo value \( \hat{S}_i = n \hat{S} - (n - 1)\hat{S}_{-i} \), where \( \hat{S} \) is the species richness from all ‘\( n \)’ sample units.

**Step 4:** Calculate first order Jackknife estimator of species richness as

\[ S_{jack}^* = \frac{1}{n} \sum_{i=1}^{n} \hat{S}_i \]

**Note:** In step 1, if two observations are deleted then we have second order jackknife estimator of species richness and so on.

Depending on the number of unique species in the deleted observations, Smith and van Belle [7] described a closed form of the first order jackknife estimator of species richness as:
where, \( r_i \) is the number of species that are found only in \( i^{th} \) sample unit, also

\[
\sum_{i=1}^{n} r_i = L
\]
as given in notation, using \( L \) we have the jackknife estimator as

\[
S_{jack}^* = S_{obs} + \frac{n-1}{n} L
\]

The standard error of jackknife estimator is given by

\[
SE = \sqrt{\frac{\sum_{i=1}^{n}(\bar{S}_i - S_{jackk})^2}{n(n-1)}}
\]

where, \( \bar{S}_i \), is the pseudo value of \( i^{th} \) sample, \( S_{jackk}^* \) is the jackknife estimator and ‘\( n \)’ is the number of sample units.

5. **Bootstrap Estimator of Species Richness**

The bootstrap method was first proposed by Bradly Efron [8] as a resampling Statistical method for estimating the sampling distribution of an estimator by sampling with replacement from the original sample. Which offers a nonparametric means of estimating parameters, beside statistical applications is now widely used by ecologist to estimate the diversity indices and unknown population characteristics. A bootstrap estimator and its variance were found in Smith and van Belle [7] for estimating species richness in ecological study.

The bootstrap method differs from the jackknife in the mean of sampling the original sample data. The subsamples (Bootstrap samples) of observed sample used to repeatedly calculate the parameter of interest are each selected at random with replacement from the original sample. The procedure of bootstrap estimates of species richness is described in the following steps.

**Step 1:** Generate a random sample of size ‘\( n \)’ (observed number of individual) from the sampled individuals with replacement, called bootstrap sample.

**Step 2:** Find the proportion say, ‘\( p_i \)’ of the observed individual for \( i^{th} \) species in the generated bootstrap sample.
Step 3: Calculate pseudo values from bootstrap sample as

\[ S_{\text{boot}(i)} = S_{\text{obs}} + \sum_{i=1}^{s_{\text{obs}}} (1 - p_i)^n \]

Step 4: Repeatedly calculate \( S_{\text{boot}(i)} \) of step 3 a large number of times, usually 20-200 times i.e. calculate for each Bootstrap sample (Bootstrap samples are considering usually 20 to 200 times).

Step 5: Finally calculate the bootstrap estimate of species richness as the average of the pseudo values calculated from the ‘B’ bootstrap samples i.e.

\[ S^*_{\text{boot}} = \frac{1}{B} \sum_{i=1}^{B} S_{\text{boot}(i)}. \]

Which gives the true estimate of species richness under bootstrap mechanism.

The standard error of the bootstrap estimate of species richness is given by

\[ SE = \sqrt{\frac{\sum_{i=1}^{B}(S_{\text{boot}(i)} - S^*_{\text{boot}})^2}{B - 1}} \]

where, \( S_{\text{boot}(i)} \); is the species richness in the ‘i’th bootstrap sample, \( S^*_{\text{boot}} \) is the actual bootstrap estimate and ‘B’ is the number of bootstrap samples taken.

6. NUMERICAL ILLUSTRATION

In this subsection of the study, we made an attempt to illustrate the three prominent nonparametric estimators of species richness, namely, Chao, Jackknife and Bootstrap considering arbitrarily chosen data for a particular biocommunity say, X in an area under study. Let the data relating to the biocommunity ‘X’ are supposed to be expressed in the following table.

<table>
<thead>
<tr>
<th>Sample units</th>
<th>Species found</th>
<th>Individual/ Status</th>
<th>Sample units</th>
<th>Species found</th>
<th>Individual/ Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>N1</td>
<td>S1, S4</td>
<td>3, 4</td>
<td>N19</td>
<td>--</td>
<td>----</td>
</tr>
<tr>
<td>N2</td>
<td>S2</td>
<td>1(a L)</td>
<td>N20</td>
<td>S28</td>
<td>1(a L)</td>
</tr>
<tr>
<td>N3</td>
<td>S20</td>
<td>1(a L)</td>
<td>N21</td>
<td>S17</td>
<td>3</td>
</tr>
<tr>
<td>N4</td>
<td>S16</td>
<td>1(a L)</td>
<td>N22</td>
<td>S23</td>
<td>2(b L)</td>
</tr>
<tr>
<td>N5</td>
<td>S22</td>
<td>1(a L)</td>
<td>N23</td>
<td>S9</td>
<td>1(a L)</td>
</tr>
<tr>
<td>N6</td>
<td>S12</td>
<td>2(b L)</td>
<td>N24</td>
<td>--</td>
<td>----</td>
</tr>
</tbody>
</table>
From the table we supposed to have 80 individuals observed in 30 different species from 36 sample units (sites), in which singleton species ‘a’ = 15; doubleton species ‘b’ = 5 and species unique to the sample unit ‘L’ = 20. From these data we will estimate the above three nonparametric estimators of species richness one by one.

### 7. Chao Estimate of Species Richness

\[
S_{\text{cho}}^* = S_{\text{obs}} + \frac{a^2}{2b} = 30 + \frac{15^2}{2 \times 5} = 52.5
\]

\[
SE\ of\ S_{\text{cho}}^* = \sqrt{\text{Var}(S_{\text{cho}}^*)} = b \left[ \left( \frac{R}{4} \right)^4 + R^3 + \left( \frac{R}{2} \right)^2 \right]
\]

\[
= \sqrt{5 \left[ \left( \frac{15}{4} \right)^4 + \left( \frac{15}{5} \right)^3 + \left( \frac{15}{5} \right)^2 \right]} = \sqrt{147.8320} = 12.1586
\]

### 8. Jackknife Estimate of Species Richness

\[
S_{\text{jack}}^* = \frac{1}{n} \sum_{i=1}^{n} \hat{S}_i = \frac{1}{36} [\hat{S}_1 + \hat{S}_2 + \cdots + \hat{S}_{36}]
\]

\[
= \frac{1}{36} \times 1990 = 55.28
\]
9. Bootstrap Estimate of Species Richness

One of the nice things about bootstrapping is that it can be done in EXCEL. MS-EXCEL 2007 (with ANALYSIS TOOLTIPK in MS-Office 2003) has a random number generator in the statement “=RANDBETWEEN (N1, N2)”, where N1 and N2 represent the range of the random numbers to be generated. The next set of numbers is the random samples, with replacement from the original data set, which we get by the statement $HLOOKUP(C3,$$C$1:$L$2,2,$FALSE$) technically we called bootstrap samples. Where C3, denotes a random number entry, for which we need to find the corresponding entry in the original data row, in the array $C$1:$L$2, the first row is the index value corresponding to a data entry in the next row and final entry $FALSE$ insures that the function returns the exact required value.

The bootstrap estimate of species richness calculated for the arbitrarily chosen data is

$$S_{boot}^* = \frac{1}{B} \sum_{i=1}^{B} S_{boot(i)} = \frac{1}{10} [399.6026] = 39.96$$

$$SE \ of \ S_{boot}^* = \sqrt{\frac{\sum_{i=1}^{B} (S_{boot(i)} - S_{boot}^*)^2}{B-1}} = \sqrt{1.59602444} = 1.2633$$

10. Discussion

Theoretically, the total species richness can be determined for any biocommunity because number of species is limited. But in practice, species counting of a particular biocommunity for example, spiders face the number of operational disadvantages or completion of complete enumeration of species often requires extraordinary efforts and is almost unattainable goal. For such disadvantages of counting total number of species, Ecologist employed the Statistical estimators to estimate species richness of a community in an area. Bunge and Fitzpatrick [9], Colwell and Coddington, [10], and Walther et al. [11] etc. have studied on estimating species richness extensively. According to Walther and Morand [12], in ecological contexts, three types of method have been used to estimate species richness namely, the fitting of species- abundance distributions, the extrapolation of species accumulation curves and non-parametric
estimators. This study covers the third method of species richness in which we have discussed the three prominent nonparametric estimators, namely, Chao, Jackknife and Bootstrap.

For species richness, at present, no clear consensus as to the best nonparametric approach is available. In different situations different estimators give good estimate of number of species of a community in an area or habitat, although the studies by Walther and Morand [12] on parasite and Magurran [13] on fish communities show the Chao estimator used with presence-absence data to perform best. Also the study of Hellmann and Fowler [14] compared the jackknife and bootstrap methods and determined that where sample size was small the jackknife behave best but statistically bootstrap perform better on the basis of re-sampling the few samples available for the population.

11. CONCLUSION

Here we have observed that the Chao estimates 52.5 species with a standard error of 12.1586, Jackknife estimates 55.28 species with a standard error of 2.9943 and the Bootstrap estimates 39.96 species with a standard error of 1.2633. Hence the standard errors of these three nonparametric estimators of species richness argued that bootstrap is the most suitable species richness estimator with minimum standard error in Ecological study. Thus we conclude with the argument that the bootstrap estimator performs reasonably well in low intensity sampling in species rich and species poor populations which has a usual tendency to underestimate the actual variances.

REFERENCE


Chavan Kr. Sarmah
Assistant Professor,
Department of Statistics,
B.H. College, Howly, Barpeta, Assam, India.
E-Mail: sarmahchavan@rediffmail.com