Diversity indices to test the goodness of fit to the broken stick distribution

David Almorza¹, María Hortensia García² and Juan Carlos Salerno³

¹ University of Cádiz, Spain
² University of Salamanca, Spain
³ Genetic Institute Ewald Favret, INTA Castelar, Buenos Aires, Argentina.

Abstract
In this paper, we present the Shannon diversity index, the Shannon exponential index and the Margalef diversity index to test the goodness of fit to the broken stick distribution in several populations. The chi-squared test is the most common test to fit the broken stick distribution, but it has several problems. With an example, we show a situation in which a test based on diversity indices improves the results obtained by using the chi-square test.

Keywords: broken stick, diversity indices, goodness of fit.

1. Introduction

The broken stick model was proposed by MacArthur [9] and it was soon extended to model the abundance of species in a habitat.

Recent works (Delport et al. [4]; Harris et al. [5]) have shown that this model is also applied to new fields of interest by geneticists and botanists.

To study the goodness of fit of the broken stick model, the chi-squared test is usually applied (Magurran [10]), although several authors have been critical of these applications (Hughes [7]; Lambshead and Platt [8]).

Almorza and Peinado [1] proposed a different test using the inverse Simpson diversity index to be applied in studies with two or more populations, and they extended (Almorza et al. [2]) this result to the Simpson diversity index.

Hill [6] used the Shannon exponential index as an index of diversity. Another of the most widely used indices is the Margalef diversity index (Margalef [11]), which is also used in a study on diversity of plankton (De León and Chalar [3]).

In this work, we extend this result to the Shannon diversity index, the Shannon exponential index and the Margalef diversity index and, in this way, we complete with the main diversity indices.

2. Materials and Methods

Let us consider a habitat occupied by \( N \) individuals from \( S \) different species. \( N_i \) will denote the number of individuals of the species \( i \) \( (i = 1, 2, \ldots, S) \), and species will be ordered in an ascendant way as a function of \( N_i \) \( (N_1 \leq N_2 \leq \ldots \leq N_s \) where \( N_1 + N_2 + \ldots + N_S = N) \).

The estimation of \( N_i \) by the broken stick model is

\[
N_i = N \left[ \frac{1}{S} + \frac{1}{S-1} + \frac{1}{S-2} + \ldots + \frac{1}{S-i+1} \right]
\]

\( i = 1, 2, \ldots, S. \)

In this way, the probability that an individual is of the species \( i \) in the study habitat is:

\[
p_i = \frac{1}{S} \left[ \frac{1}{S} + \frac{1}{S-1} + \frac{1}{S-2} + \ldots + \frac{1}{S-i+1} \right], \text{ where } 0 \leq p_i \leq 1, \forall \ i = 1, 2, \ldots, S, \text{ and } \sum_{i=1}^{S} p_i = 1 \text{ for } p_i \leq p_j \forall \ i \leq j, \text{ where } i, j = 1, 2, \ldots, S.
\]

The inverse Simpson diversity index, the Simpson diversity index, the Shannon diversity index, the Shannon exponential index and the Margalef diversity index are defined, respectively, by:

\[
D = \frac{1}{S} \sum_{i=1}^{S} p_i^2 \quad \text{and} \quad 1 \leq D \leq S; \quad D = 1 - \sum_{i=1}^{S} p_i^2
\]

\[
H = -\sum_{i=1}^{S} p_i \ln p_i \quad \text{and} \quad 0 \leq D \leq \frac{S-1}{S}; \quad H = \sum_{i=1}^{S} p_i \ln p_i
\]
If two populations with species $S_1$ and $S_2$ ($S_1 > S_2$), respectively, are modeled by the broken stick model, then:

a) $D'(S_1) > D'(S_2)$

b) $D(S_1) > D(S_2)$

c) $H(S_1) > H(S_2)$

d) $e^{H(S_1)} > e^{H(S_2)}$

e) $D_{mg}(S_1) > D_{mg}(S_2)$

**Proof**

a) Almorza and Peinado [1]

b) Almorza et al. [2]

c) We consider $S_1$ and $S_2 = S_1 - 1$. We obtain the Shannon diversity index for both values. Then, we obtain that:

For $S_1$, the Shannon diversity index is:

$$H(S_1) = -\sum_{i=1}^{S_1} \frac{1}{S_1} \left( \frac{1}{S_1} + \frac{1}{S_1-1} + \ldots + \frac{1}{S_1-i+1} \right)$$

$$= -\sum_{i=1}^{S_1} \frac{1}{S_1} \left( \frac{1}{S_i} \right) + \frac{1}{S_i} - \frac{1}{S_i} + \frac{1}{S_i} - \frac{1}{S_i} + \ldots + \frac{1}{S_i} - \frac{1}{S_i} + \frac{1}{S_i} - \frac{1}{S_i} + \frac{1}{S_i} - \frac{1}{S_i} + \ldots + \frac{1}{S_i} - \frac{1}{S_i} + \frac{1}{S_i} - \frac{1}{S_i} + \ldots + \frac{1}{S_i} - \frac{1}{S_i} + \frac{1}{S_i}$$

$$= -\ln \left( \frac{1}{S_1} \right) \frac{1}{S_1} \left( \frac{1}{S_1} \right) + \frac{1}{S_1} - \frac{1}{S_1} + \frac{1}{S_1} - \frac{1}{S_1} + \ldots + \frac{1}{S_1} - \frac{1}{S_1} + \frac{1}{S_1} - \frac{1}{S_1} + \ldots + \frac{1}{S_1} - \frac{1}{S_1} + \frac{1}{S_1}$$

$$= -\ln \left( \frac{1}{S_1} \right) \frac{1}{S_1} \left( \frac{1}{S_1} \right) + \frac{1}{S_1} - \frac{1}{S_1} + \frac{1}{S_1} - \frac{1}{S_1} + \ldots + \frac{1}{S_1} - \frac{1}{S_1} + \frac{1}{S_1} - \frac{1}{S_1} + \ldots + \frac{1}{S_1} - \frac{1}{S_1} + \frac{1}{S_1}$$

$$= -\ln \left( \frac{1}{S_1} \right) \frac{1}{S_1} \left( \frac{1}{S_1} \right) + \frac{1}{S_1} - \frac{1}{S_1} + \frac{1}{S_1} - \frac{1}{S_1} + \ldots + \frac{1}{S_1} - \frac{1}{S_1} + \frac{1}{S_1} - \frac{1}{S_1} + \ldots + \frac{1}{S_1} - \frac{1}{S_1} + \frac{1}{S_1}$$

$$= -\ln \left( \frac{1}{S_1} \right) \frac{1}{S_1} \left( \frac{1}{S_1} \right) + \frac{1}{S_1} - \frac{1}{S_1} + \frac{1}{S_1} - \frac{1}{S_1} + \ldots + \frac{1}{S_1} - \frac{1}{S_1} + \frac{1}{S_1} - \frac{1}{S_1} + \ldots + \frac{1}{S_1} - \frac{1}{S_1} + \frac{1}{S_1}$$

$$= -\ln \left( \frac{1}{S_1} \right) \frac{1}{S_1} \left( \frac{1}{S_1} \right) + \frac{1}{S_1} - \frac{1}{S_1} + \frac{1}{S_1} - \frac{1}{S_1} + \ldots + \frac{1}{S_1} - \frac{1}{S_1} + \frac{1}{S_1} - \frac{1}{S_1} + \ldots + \frac{1}{S_1} - \frac{1}{S_1} + \frac{1}{S_1}$$

...
For $S_2 = S_1 - 1$, the Shannon diversity index is:

$$H_{(S_2)} = -\sum_{i=1}^{S_2} \frac{1}{S_2} \left( \frac{1}{S_1} + \frac{1}{S_2} - 1 + \ldots + \frac{1}{S_2^{i-1}} \right)^{\frac{1}{S_2}}$$

$$= \ln \left( \frac{1}{S_1} \right) \left( \frac{1}{S_1} + \frac{1}{S_2} - 1 + \ldots + \frac{1}{S_2^{i-1}} \right)^{\frac{1}{S_2}}$$

$$= \ln \left( \frac{1}{S_1} \right) \left( \frac{1}{S_1} + \frac{1}{S_2} - 1 + \ldots + \frac{1}{S_2^{i-1}} \right)^{\frac{1}{S_2}} = -\ln A \quad (3.1)$$

We have:

- $S_1 > S_2 > S_2 - 1 > \ldots > 2 > 1$
- $\frac{1}{S_1} < \frac{1}{S_2} - 1 < \frac{1}{S_2 - 2} < \ldots < \frac{1}{2}$

Comparing similar terms in (3.1) and (3.2), we obtain that: $A < B$.

From applying logarithmic properties, we conclude: $H_{(S_1)} > H_{(S_2)}$.

d) Based on previous results and taking into account the properties of the exponential function, we obtain: $e^{H_{(S_1)}} > e^{H_{(S_2)}}$.

e) Given that $S_1 > S_2$, then $S_1 - 1 > S_2 - 1$.

Dividing both members of the inequality for $\ln N$, we have: $D_{mg}(S_1) > D_{mg}(S_2)$. 

Copyright (c) 2014 International Journal of Computer Science Issues. All Rights Reserved.
4. Application

As an example of the importance of these results in the adjustment to the broken stick model, we have developed the following situation. The information is artificial, but it is useful to illustrate the theoretical results.

The chi-squared test is most often used to measure the goodness of fit of the broken stick model (Magurran [10]). This method has been criticized in various ways by different authors, including Hughes [7] and Lambshead and Platt [8] among others.

We consider two habitats with \( S_1 = 7 \) and \( S_2 = 8 \), as shown in Table 1.

Table 1: Information used to fit the broken stick model in two populations.

<table>
<thead>
<tr>
<th>Habitat 1</th>
<th>( S = 7 )</th>
<th>Habitat 2</th>
<th>( S = 8 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species</td>
<td>Individuals</td>
<td>Species</td>
<td>Individuals</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>11</td>
<td>4</td>
<td>14</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td>5</td>
<td>21</td>
</tr>
<tr>
<td>6</td>
<td>27</td>
<td>6</td>
<td>36</td>
</tr>
<tr>
<td>7</td>
<td>37</td>
<td>7</td>
<td>77</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td></td>
<td>143</td>
</tr>
</tbody>
</table>

Using the chi-squared test, we found that both habitats are compatible with the broken stick model.

The problem is that habitat 2 was obtained from a simulation of a geometric model, and the test cannot find significant differences between this model and the broken stick model. Habitat 1 was obtained from a simulation model of the broken stick (in both cases, we used the Species Diversity and Richness software version 4.0 [12]). This aspect, which was not detected by the chi-squared test, is revealed by the application of the results, as shown in Table 2.

Table 2: Values of the measures of diversity for the two collections of data.

<table>
<thead>
<tr>
<th></th>
<th>Habitat 1</th>
<th>Habitat 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of species</td>
<td>( S_1 = 7 )</td>
<td>( S_2 = 8 )</td>
</tr>
<tr>
<td>The Inverse Simpson diversity index</td>
<td>( D'_1 = 4.224 )</td>
<td>( D'_2 = 3.199 )</td>
</tr>
<tr>
<td>The Shannon diversity index</td>
<td>( H_1 = 1.595 )</td>
<td>( H_2 = 1.423 )</td>
</tr>
</tbody>
</table>

In the cases of inverse Simpson diversity index, the Shannon diversity index, the Shannon exponential index and the Margalef diversity index it is verified that:

\[
D'(S_1) < D'(S_2); \quad H(S_1) < H(S_2); \quad e^{H(S_1)} < e^{H(S_2)}; \quad D_{mg}(S_1) < D_{mg}(S_2)
\]

However, because \( S_1 > S_2 \), it indicates (by the previous results) that there is a failure of the fit to the broken stick model, as already stated.

5. Conclusion

We showed, with this example, that a test based on diversity indices improves the results obtained by using the chi-square test.

References


David Almorza. Dr. Department of Statistics and Operational Research. Faculty of Labour Sciences. University of Cádiz (Spain). Vice-Rector of Social Responsibility and University Services.

María Hortensia García. Department of Statistics. University of Salamanca (Spain).