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

David Almorza[1], María Hortensia García[2] and Juan Carlos Salerno[3]

[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 critics 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,…, S), and species will be ordered in an ascendant way as a function of N_i ($N_1 \le N_2 \le ..... \le N_S$ where $N_1 + N_2 + ... + N_S = N$).

The estimation of N_i by the broken stick model is

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

i = 1,2,...., S.

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

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

$$D' = \frac{1}{\sum_{i=1}^{S} p_i^2} \text{ and } 1 \le D' \le S \, ; \quad D = 1 - \sum_{i=1}^{S} p_i^2$$

and $0 \le D \le \frac{S-1}{S}$; $H = -\sum_{i=1}^{S} p_i \ln p_i$ and

$0 \le H \le \ln S$; $I_{ex}=e^{H}$ and $1 \le I_{ex} \le e^{\log S}$;
$D_{mg} \le \dfrac{S-1}{\ln N}$ and $0 \le D_{mg} \le \dfrac{N-1}{\ln N}$.

3. Results

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:

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

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

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

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

$$=-\ln\left(\frac{1}{S_1}\right)^{\frac{1}{S_1}\frac{1}{S_1}}\cdot\left(\frac{1}{S_1}\right)^{\frac{1}{S_1}\frac{1}{S_1}}$$

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

$$-\ln\left(\frac{1}{S_1}\right)^{\frac{1}{S_1}\left(\frac{1}{S_1}+\frac{1}{S_1-1}+\ldots+\frac{1}{2}\right)}\cdot$$

$$\left(\frac{1}{S_1}+\frac{1}{S_1-1}+\ldots+\frac{1}{2}\right)^{\frac{1}{S_1}\left(\frac{1}{S_1}+\frac{1}{S_1-1}+\ldots+\frac{1}{2}\right)}-$$

$$-\ln\left(\frac{1}{S_1}\right)^{\frac{1}{S_1}\left(\frac{1}{S_1}+\frac{1}{S_1-1}+\ldots+\frac{1}{2}+1\right)}\cdot$$

$$\left(\frac{1}{S_1}+\frac{1}{S_1-1}+\ldots+\frac{1}{2}+1\right)^{\frac{1}{S_1}\left(\frac{1}{S_1}+\frac{1}{S_1-1}+\ldots+\frac{1}{2}+1\right)}=$$

$$=-\ln\left(\frac{1}{S_1}\right)^{\frac{1}{{S_1}^2}}\cdot\left(\frac{1}{S_1}\right)^{\frac{1}{{S_1}^2}}\cdot\left(\frac{1}{S_1}\right)^{\frac{1}{S_1}\left(\frac{1}{S_1}+\frac{1}{S_1-1}\right)}\cdot$$

$$\left(\frac{1}{S_1}+\frac{1}{S_1-1}\right)^{\frac{1}{S_1}\left(\frac{1}{S_1}+\frac{1}{S_1-1}\right)}\cdot\ldots\cdot\left(\frac{1}{S_1}\right)^{\frac{1}{S_1}\left(\frac{1}{S_1}+\frac{1}{S_1-1}+\ldots+\frac{1}{2}\right)}\cdot$$

$$\cdot\left(\frac{1}{S_1}+\frac{1}{S_1-1}+\ldots+\frac{1}{2}\right)^{\frac{1}{S_1}\left(\frac{1}{S_1}+\frac{1}{S-1}+\ldots+\frac{1}{2}\right)}\cdot$$

$$\left(\frac{1}{S_1}\right)^{\frac{1}{S_1}\left(\frac{1}{S_1}+\frac{1}{S_1-1}+\ldots+\frac{1}{2}+1\right)}\cdot$$

$$\cdot\left(\frac{1}{S_1}+\frac{1}{S_1-1}+\ldots+\frac{1}{2}+1\right)^{\frac{1}{S_1}\left(\frac{1}{S_1}+\frac{1}{S_1-1}+\ldots+\frac{1}{2}+1\right)}=$$

$$=-\ln\left(\frac{1}{S_1}\right)^{\frac{1}{S_1^2}}\left(\frac{1}{S_1}\right)^{\frac{1}{S_1^2}}\left(\frac{1}{S_1}\right)^{\frac{1}{S_1^2}}\left(\frac{1}{S_1}\right)^{\frac{1}{S_1}\cdot\frac{1}{S_1-1}}\cdot$$

$$\left(\frac{1}{S_1}+\frac{1}{S_1-1}\right)^{\frac{1}{S_1^2}}\left(\frac{1}{S_1}+\frac{1}{S_1-1}\right)^{\frac{1}{S_1^2}\cdot\frac{1}{S_1-1}}\ldots$$

$$\left(\frac{1}{S_1}\right)^{\frac{1}{S_1^2}}\left(\frac{1}{S_1}\right)^{\frac{1}{S_1}\cdot\frac{1}{S_1-1}}\ldots\left(\frac{1}{S_1}\right)^{\frac{1}{S_1}\cdot\frac{1}{2}}\cdot$$

$$\left(\frac{1}{S_1}+\frac{1}{S_1-1}+\ldots+\frac{1}{2}\right)^{\frac{1}{S_1^2}}\left(\frac{1}{S_1}+\frac{1}{S_1-1}+\ldots+\frac{1}{2}\right)^{\frac{1}{S_1}\cdot\frac{1}{S_1-1}}$$

$$\ldots\left(\frac{1}{S_1}+\frac{1}{S_1-1}+\ldots+\frac{1}{2}\right)^{\frac{1}{S_1}\cdot\frac{1}{2}}\cdot$$

$$\left(\frac{1}{S_1}\right)^{\frac{1}{S_1^2}}\left(\frac{1}{S_1}\right)^{\frac{1}{S_1}\cdot\frac{1}{S_1-1}}\ldots\left(\frac{1}{S_1}\right)^{\frac{1}{S_1}\cdot\frac{1}{2}}\left(\frac{1}{S_1}\right)^{\frac{1}{S_1}\cdot 1}$$

$$\left(\frac{1}{S_1}+\frac{1}{S_1-1}+\ldots+\frac{1}{2}+1\right)^{\frac{1}{S_1^2}}\cdot$$

$$\left(\frac{1}{S_1}+\frac{1}{S_1-1}+\ldots+\frac{1}{2}+1\right)^{\frac{1}{S_1}\cdot\frac{1}{S_1-1}}\ldots$$

$$\left(\frac{1}{S_1}+\frac{1}{S_1-1}+\ldots+\frac{1}{2}+1\right)^{\frac{1}{S_1}\cdot\frac{1}{2}}\cdot$$

$$\left(\frac{1}{S_1}+\frac{1}{S_1-1}+\ldots+\frac{1}{2}+1\right)^{\frac{1}{S_1}\cdot 1}=$$

$$=-\ln\left(\frac{1}{S_1}\right)^{\frac{1}{S_1^2}\cdot S_1}\left(\frac{1}{S_1}\right)^{\frac{1}{S_1^2}}\left(\frac{1}{S_1}\right)^{\frac{1}{S_1}\cdot\frac{1}{S_1-1}\cdot(S_1-1)}\cdot$$

$$\left(\frac{1}{S_1}+\frac{1}{S_1-1}\right)^{\frac{1}{S_1}\left(\frac{1}{S_1}+\frac{1}{S_1-1}\right)}\ldots\left(\frac{1}{S_1}\right)^{\frac{1}{S_1}\cdot\frac{1}{2}\cdot 2}$$

$$\left(\frac{1}{S_1}+\frac{1}{S_1-1}+\ldots+\frac{1}{2}\right)^{\frac{1}{S_1}\left(\frac{1}{S_1}+\frac{1}{S_1-1}+\ldots+\frac{1}{2}\right)}\cdot$$

$$\left(\frac{1}{S_1}\right)^{\frac{1}{S_1}\cdot\frac{1}{1}\cdot 1}$$

$$\left(\frac{1}{S_1}+\frac{1}{S_1-1}+\ldots+\frac{1}{2}+1\right)^{\frac{1}{S_1}\left(\frac{1}{S_1}+\frac{1}{S_1-1}+\ldots+\frac{1}{2}+1\right)}=$$

$$=-\ln\left(\frac{1}{S_1}\right)^{\frac{S_1}{S_1}}\left(\frac{1}{S_1}\right)^{\frac{1}{S_1^2}}\cdot$$

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

$$\left(\frac{1}{S_1}+\frac{1}{S_1-1}+\ldots+\frac{1}{2}\right)^{\frac{1}{S_1}\left(\frac{1}{S_1}+\frac{1}{S_1-1}+\ldots+\frac{1}{2}\right)}\cdot$$

$$\left(\frac{1}{S_1}+\frac{1}{S_1-1}+\ldots+\frac{1}{2}+1\right)^{\frac{1}{S_1}\left(\frac{1}{S_1}+\frac{1}{S_1-1}+\ldots+\frac{1}{2}+1\right)}=$$

$$-\ln\left(\frac{1}{S_1}\right)^{\frac{S_1^2+1}{S_1^2}}\left(\frac{1}{S_1}+\frac{1}{S_1-1}\right)^{\frac{1}{S_1}\left(\frac{1}{S_1}+\frac{1}{S_1-1}\right)}\ldots$$

$$\left(\frac{1}{S_1}+\frac{1}{S_1-1}+\ldots+\frac{1}{2}\right)^{\frac{1}{S_1}\left(\frac{1}{S_1}+\frac{1}{S_1-1}+\ldots+\frac{1}{2}\right)}$$

$$\left(\frac{1}{S_1}+\frac{1}{S_1-1}+\ldots+\frac{1}{2}+1\right)^{\frac{1}{S_1}\left(\frac{1}{S_1}+\frac{1}{S_1-1}+\ldots+\frac{1}{2}+1\right)}=$$

$$=-\ln A \qquad (3.1)$$

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

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

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

$$=-\sum_{i=1}^{S_1-1}\ln\left(\frac{1}{S_1-1}\right)^{\frac{1}{S_1-1}\left(\frac{1}{S_1-1}+\frac{1}{S_1-2}+\ldots+\frac{1}{S_1-i}\right)}\cdot$$

$$\left(\frac{1}{S_1-1}+\frac{1}{S_1-2}+\ldots+\frac{1}{S_1-i}\right)^{\frac{1}{S_1-1}\left(\frac{1}{S_1-1}+\frac{1}{S_1-2}+\ldots+\frac{1}{S_1-i}\right)}=$$

$$=-\ln\left(\frac{1}{S_1-1}\right)^{\frac{(S_1-1)^2+1}{(S_1-1)^2}}\cdot$$

$$\left(\frac{1}{S_1-1}+\frac{1}{S_1-2}\right)^{\frac{1}{S_1-1}\left(\frac{1}{S_1-1}+\frac{1}{S_1-2}\right)}\cdot$$

$$\left(\frac{1}{S_1-1}+\frac{1}{S_1-2}+\ldots+\frac{1}{2}\right)^{\frac{1}{S_1-1}\left(\frac{1}{S_1-1}+\frac{1}{S_1-2}+\ldots+\frac{1}{2}\right)}\cdot$$

$$\left(\frac{1}{S_1-1}+\frac{1}{S_1-2}+\ldots+\frac{1}{2}+1\right)^{\frac{1}{S_1-1}\left(\frac{1}{S_1-1}+\frac{1}{S_1-2}+\ldots+\frac{1}{2}+1\right)}=$$

$$= - \ln B \qquad (3.2)$$

We have:

- $S_1 > S_1 - 1 > S_1 - 2 > \ldots\ldots\ldots. > 2 > 1$
- $\frac{1}{S_1} < \frac{1}{S_1-1} < \frac{1}{S_1-2} < \ldots\ldots\ldots < \frac{1}{2} < 1$
- $\frac{1}{S_1^2} < \frac{1}{(S_1-1)^2} < \frac{1}{(S_1-2)^2} < \ldots\ldots\ldots < \frac{1}{2^2} < 1$

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: $D_{mg\,(S_1)} > D_{mg\,(S_2)}$.

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.

Habitat 1	***S* = 7**	**Habitat 2**	***S* = 8**
Species	Individuals	Species	Individuals
1	1	1	2
2	4	2	3
3	8	3	4
4	11	4	14
5	12	5	21
6	27	6	36
7	37	7	77
		8	143

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.

	Habitat 1	**Habitat 2**
Number of species	$S_1 = 7$	$S_2 = 8$
The Inverse Simpson diversity index	$D'_1 =$ 4.224	$D'_2 =$ 3.199
The Shannon diversity index	$H_1 =$ 1.595	$H_2 =$ 1.423
The Shannon exponential index	$I_{ex1} =$ 4.931	$I_{ex2} =$ 4.149
The Margalef diversity index	$D_{mg1} =$ 1.303	$D_{mg2} =$ 1.227

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:

$D'_{(S_1)} < D'_{(S_2)}$; $H_{(S_1)} < H_{(S_2)}$; $e^{H(S_1)} < e^{H(S_2)}$; $D_{mg\,(S_1)} < D_{mg\,(S_2)}$

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

[1] D. Almorza; A. Peinado. *Análisis de diversidad a partir del modelo del bastón roto*. Información Tecnológica **12**, 5 (2001) 145-147.

[2] D. Almorza; A. Peinado; C. Valero; R. Boggio; A. Rodríguez. *Índices de diversidad de Simpson en el modelo del bastón roto*. Jorma VII (2001) 7-8.

[3] L. De León; G. Chalar. *Abundancia y diversidad del fitoplancton en el Embalse de Salto Grande (Argentina – Uruguay). Ciclo estacional y distribución espacial*. Limnetica **22**, 1-2 (2003) 103-113.

[4] W. Delport; M. Cunningham; B. Olivier; O. Preisig; S.W. van der Merwe. *A population genetics pedigree perspective on the transmission of Helicobacter pylori*. Genetics **174** (2006) 2107–2118.

[5] K. Harris; P.K. Subudhi; A. Borrell; D. Jordan; D. Rosenow; H. Nguyen; P. Klein; R. Klein; J. Mullet. *Sorghum stay-green QTL individually reduce post-flowering drought-induced leaf senescence*. Journal of Experimental Botany **58**, 2 (2007) 327–338.

[6] M.O. Hill. *Diversity and evenness: a unifying notation and its consequences*. Ecology **54** (1973) 427-432.

[7] R.G. Hughes. *Theories and models of species abundance*. American Naturalist **128** (1986) 879-899.

[8] J. Lambshead; H.M. Platt. *Structural patterns of marine benthics assemblages*

and their relationships with empirical statistical models. Proceedings of the 19th European Marine Biology Symposium, Plymouth (1984) 371-380.

[9] R.H. MacArthur. *On the relative abundance of bird species*. Proceedings of the National Academy of Sciences of the United States of America **43** (1957) 293-295.

[10] A.E. Magurran. *Diversidad ecológica y su medición*, 1ª edición. Ediciones Vedra, Barcelona- España (1989).

[11] R. Margalef. *Information theory in ecology*. General Systematics **3** (1958) 36-71.

[12] R. M. Seaby; P. A. Henderson. *Species Diversity and Richness Version 4*. Pisces Conservation Ltd., Lymington, England (2006).

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).

Juan Carlos Salerno. Dr. Genetic Institute Ewald Favret. Castelar. Buenos Aires (Argentina).