Short Communication

A Java program to test homogeneity of samples and examine sampling completeness

WenJun Zhang

School of Life Sciences, Sun Yat-sen University, Guangzhou 510275, China; International Academy of Ecology and Environmental Sciences, Hong Kong

E-mail: zhwj@mail.sysu.edu.cn, wjzhang@iaees.org

Received 9 July 2011; *Accepted 22 August 2011*; *Published online 1 September 2011*

IAEES

Abstract

A Java program to test the homogeneity of samples and examine sampling completeness was presented in this study. The program was based on the model of Coleman et al. (1982) for random placement hypothesis and the algorithm of Zhang et al. (1999). The program was used to test samples' homogeneity and examine sampling completeness for four arthropod sampling data sets.

Keywords sample homogeneity; sampling completeness; statistic test; Java program.

1 Introduction

In food web sampling studies, we need to record all possible taxa (species, families, etc.) in the community. Enough samples should be taken to record enough taxa. To examine sampling completeness, a yield-effort curve may be drawn, which plots the cumulative number of taxa caught or observed (y-axis) against the cumulative effort of sampling (x-axis) (Cohen, 1978; Dickerson and Robinson, 1985; Cohen et al., 1993; Zhang and Schoenly, 1999). If sampling stops while the yield-effort curve is still rapidly increasing, then the community derived from this sampling is incomplete. Nevertheless, if sampling ceases when the slope of the yield-effort curve reaches zero or close to zero, the sampling is probably complete. In addition, the correlation based eco-interaction network studies require the homogeneity of samples or environment. How to ensure sample homogeneity is also a necessity work in these studies. Bias from sample order can be corrected by bootstrap procedure. However, variation in curve shape due to environmental heterogeneity remains a likely significant source of sampling error (Zhang and Schoenly, 1999). Coleman et al. (1982) developed a statistical model to test whether individuals among the samples (of definable size) obey the random placement hypothesis which assumes a lack of correlation in the location of individuals (Zhang and Schoenly, 1999). The model of Coleman et al. can test sample homogeneity and examine sampling completeness. In this study, a Java program, based on the model of Coleman et al. (1982) and the algorithm of Zhang and Schoenly (1999) was presented. Compared to the algorithm of Zhang and Schoenly (1999), it gives the conclusion for completeness of sampling and can be easily run on web browser.

2 Algorithm

Under the random placement hypothesis, consider a collection C of N individuals from S taxa, with ni individuals in C belonging to the i-th taxon, and suppose that each member of C occurs in one of k non-

overlapping samples that have areas a_1, a_2, …, a_k. The number of taxa, s, in a given region is a random variable whose magnitude depends on the area a of the region, and the relative area is defined as $\alpha = a/\sum a_i$. The mean number of taxa, s, and the variance σ^2 are calculated as follows:

$$s(\alpha)=S-\sum(1-\alpha)^{ni},$$
$$\sigma^2(\alpha)=\sum(1-\alpha)^{ni}-\sum(1-\alpha)^{2ni}.$$

The method to test sample homogeneity is to compare the observed mean taxa richness *vs.* the sample size with the expected taxa richness *vs.* the sample size curve (Zhang and Schoenly, 1999). If 95% of the plotted points (means) of the observed curve fall two standard deviations outside the expected curve, then the observed samples are statistically more heterogeneous in taxa composition (at the 0.05 level) than sampling error (alone) can account for (Coleman et al., 1982). Thus we can conclude that these samples are more heterogeneous in taxonomic composition than is expected under the random placement hypothesis.

Bootstrap procedures are used to produce the taxa richness *vs.* the sample size curves. The curves plot the cumulative number of taxa, defined as the sum of the number of taxa in the previous sample(s) and the number of taxa in the present sample that were not observed in any previous sample. For the first sample, the cumulative number of taxa is defined to equal number of taxa found in this sample.

If random placement hypothesis is met, the samples are homogeneous, or else they are heterogeneous. If the difference of the number of taxa between the last two (cumulative) sample sizes is less than desired percent threshold, then most of the taxa are considered to be recorded and the sample size is enough.

The algorithm is implemented as a Java program, SampHomoTest, based on JDK 1.1.8, in which several classes and an HTML file is included (http://www.iaees.org/publications/software/index.asp). In sampling data file, the first row is sample ID numbers and the first column is taxon ID numbers.

3 Application

We obtained a set of arthropod data investigated in rice fields of Guangzhou, China (Data set 1: 35 samples; 19 families; Data set 2: 54 samples; 23 families; Data set 3: 60 samples; 23 families; Data set 4: 60 samples; 27 families), investigated in 2006 (Zhou, 2007).

Choose 1000 randomizations and set the sampling completeness as 0.01 (the difference of the number of taxa between the last two (cumulative) sample sizes is less than 1%). The results from the algorithm above showed that four arthropod communities are all environmentally homogeneous (all observed points fell inside the confidence interval) and the sampling is complete for the four arthropod communities (difference is around 0.5%). The results for a data set (35 samples, 19 families) are listed in Table 1.

Table 1 Test results for a data set

Sample Size	Mean Observed Number of Taxa (ONT)	Expected Number of Taxa (ENT)	Standard Devi. of Expected Number of Taxa	Lower Limit of ENT	Upper Limit of ENT	Lower Limit of ONT	Upper Limit of ONT
1	5.504	6.799	1.311	4.175	9.423	2.067	8.94
2	7.71	8.561	1.464	5.632	11.49	4.253	11.166
3	9.318	9.812	1.534	6.743	12.882	5.906	12.729
4	10.534	10.81	1.557	7.694	13.926	7.252	13.815
5	11.536	11.637	1.555	8.526	14.748	8.218	14.853
6	12.322	12.338	1.538	9.261	15.416	9.227	15.416

7	12.818	12.942	1.514	9.913	15.972	9.79	15.845
8	13.486	13.469	1.486	10.496	16.442	10.45	16.521
9	13.982	13.933	1.456	11.02	16.847	11.061	16.902
10	14.416	14.346	1.426	11.493	17.199	11.581	17.25
11	14.815	14.716	1.395	11.924	17.508	12.123	17.506
12	15.011	15.051	1.365	12.319	17.782	12.307	17.714
13	15.39	15.355	1.336	12.682	18.028	12.643	18.136
14	15.634	15.634	1.307	13.019	18.248	12.904	18.363
15	15.961	15.89	1.278	13.333	18.447	13.326	18.595
16	16.181	16.128	1.249	13.628	18.628	13.525	18.838
17	16.366	16.35	1.221	13.907	18.793	13.672	19.059
18	16.637	16.557	1.192	14.172	18.943	14.075	19.198
19	16.814	16.753	1.163	14.425	19.08	14.304	19.323
20	16.969	16.937	1.133	14.669	19.204	14.513	19.424
21	17.118	17.112	1.103	14.906	19.318	14.58	19.655
22	17.293	17.278	1.071	15.136	19.42	14.884	19.701
23	17.441	17.437	1.037	15.362	19.512	15.089	19.792
24	17.671	17.59	1.001	15.586	19.593	15.442	19.899
25	17.761	17.736	0.964	15.808	19.664	15.549	19.972
26	17.876	17.878	0.923	16.031	19.724	15.813	19.938
27	18.05	18.015	0.879	16.257	19.773	15.996	20.103
28	18.142	18.148	0.83	16.487	19.809	16.089	20.194
29	18.294	18.277	0.776	16.724	19.831	16.462	20.125
30	18.444	18.404	0.716	16.971	19.837	16.795	20.092
31	18.51	18.527	0.647	17.232	19.823	16.949	20.07
32	18.661	18.649	0.567	17.515	19.783	17.327	19.994
33	18.748	18.767	0.467	17.832	19.703	17.597	19.898
34	18.898	18.884	0.334	18.215	19.553	18.139	19.656
35	19.0	19.0	0.0	19.0	19.0	19.0	19.0

References

Cohen JE. 1978. Food Webs and Niche Space. Monographs in Population Biology 11. Princeton University Press, Princeton, NJ, USA

Cohen JE, et al. 1993. Improving food webs. Ecology, 74: 252-258

Coleman BD, Mares MA, Willig MR, et al. 1982. Randomness, area, and species richness. Ecology, 63: 1121-1133

Dickerson JE, Robinson JV. 1985. Microcosms as islands: a test of the MacArthur-Wilson equilibrium theory. Ecology, 66: 966-980

Zhang WJ, Schoenly KG. 1999. IRRI Biodiversity Software Series. II. COLLECT1 and COLLECT2: Programs for Calculating Statistics of Collectors' Curves. IRRI Technical Bulletin No.2. International Rice Research Institute, Manila, Philippines

Zhang WJ. 2011. Constructing ecological interaction networks by correlation analysis: hints from community sampling. Network Biology, 1(2): 81-98

Zhou WG. 2007. A Field Survey on Paddy Rice Arthropod Biodiversity in Northern Guangzhou. Master Degree Dissertation. Sun Yat-sen University, China