

Rarefaction Diversity: a Case Study of Polychaete Communities Using an Amended FORTRAN Program

Hwey-Lian Hsieh^{1,*} and Lung-An Li²

¹Institute of Zoology, Academia Sinica, Taipei, Taiwan 115, R.O.C.

Fax: 886-2-27858059. E-mail: zohl@ccvax.sinica.edu.tw

²Institute of Statistical Science, Academia Sinica, Taipei, Taiwan 115, R.O.C.

(Accepted September 19, 1997)

Hwey-Lian Hsieh and Lung-An Li (1998) Rarefaction Diversity: a case study of polychaete communities using an amended FORTRAN program. *Zoological Studies* 37(1): 13-21. This study presents an amended Simberloff rarefaction FORTRAN program and a case study comparing the species diversities of polychaete communities using this new program. A reexamination of the rarefaction formula revealed that the Simberloff program could be simplified. The modified program requires substantially less computer memory, and concomitantly, gains greater efficiency in calculating the expected numbers of species and their variances. For the case study, the polychaete communities in 2 regions (north and south) located in the subtidal areas off the coast of Tainan County in southwestern Taiwan were examined in September 1994 and in April 1995. A total of 724 individuals of 56 polychaete species were collected from the north region and 378 individuals of 29 species were collected from the south region. The rarefaction curves of the 2 regions were well separated. The number of species expected in the north region was greater than in the south region when the abundances of the 2 regions were rarefied to the same numbers of individuals. The rarefaction curve was steeper in the north region than in the south region, indicating a more even distribution of polychaete individuals among species in the north than in the south. In addition, the rarefaction diversity and conventional Shannon diversity and evenness index were also compared to show the significant and informative strength of the employed rarefaction method.

Key words: Species diversity, Rarefaction program, Benthic communities.

A large number of species coexisting in a given region has led to extensive studies being performed to explain the patterns of species diversity in both marine and terrestrial ecosystems (e.g., MacArthur 1965, Pianka 1966). In recent decades, the destruction of natural habitats and the introduction of exotic plants and animals into natural environments caused by human activities have drastically and rapidly reduced species diversity (e.g., Wilson 1994). Adequate measurement of species diversity is essential for understanding the control mechanisms of species diversity and the structure and function of ecosystems.

Species diversity has 2 components, species richness and species evenness; the former describes the number of species, and the latter, the relative abundance among the species in the com-

munity. The Shannon diversity index (H') and evenness index (J') (Pielou 1966a) have been extensively employed in studying the ecology of communities (e.g., Ludwig and Reynolds 1988). However, the Shannon diversity index incorporates both species richness and evenness into a single value, and hence, confounds a number of parameters that characterize community structures. These parameters include (1) the number of species (species richness), (2) relative species abundance (evenness), (3) the number of individuals sampled, and (4) the homogeneity and size of the sampling area (e.g., Smith and Grassle 1977, James and Rathbun 1981, Ludwig and Reynolds 1988). The latter 2 parameters reveal that the estimate of the Shannon index is sample-size dependent; thus, the inaccuracy of the estimation may be large, particularly

*To whom correspondence and reprint requests should be addressed.

when the sample size is small (Smith and Grassle 1977). In addition, large samples would have more species than small ones; hence, the values of the indices can not be compared, even if the collections are drawn from the same community (Heck et al. 1975, James and Rathbun 1981).

Rarefaction diversity measurement provides an alternative that avoids these difficulties by calculating species richness by scaling down all collections to the same sample size (Hurlbert 1971, Heck et al. 1975). Rarefaction has been considered an appropriate tool for defining community structure and has been used in comparing species richness among communities in various ecosystems (e.g., in avian communities, James and Rathbun 1981; in deep-sea benthos, Grassle and Maciolek 1992).

Rarefaction is a statistical method for estimating the number of species expected ($E(S_n)$) to be present in a random sample of individuals taken from any given collection. Given the number of individuals of each species in the original collection, one can calculate a series that reflects the numbers of species present in each randomly and successively drawn smaller subset of the original collection. The method then allows for the generation of a rarefaction curve. This method estimates not only the parameter of species richness, but also the confidence limits for this parameter (Heck et al. 1975); thus, communities with different species richness can be compared statistically. In addition, the shape of the curves is a graphic display of accumulation rates of relative abundance; therefore, the evenness of communities can be compared by examining the steepness of the curves and their intersection (Simberloff 1978, James and Rathbun 1981). In general, the steeper the rarefaction curve is, the higher the evenness.

A rarefaction program written in FORTRAN IV was developed by Simberloff (1978) and slightly modified by Krebs (1989) later. This program needs to store 6000 or more intermediate terms if the total number of individuals is larger than 6000. These terms are used to compute the expected numbers of species present in each random sample and their variances. In other words, this program requires a substantial amount of memory for storage. As a result, numerical overflow problems often occur, resulting in the termination of the execution of the program. After reexamining the original formula for rarefaction, we found no need for such large dimensions. The purposes of this study are to present 1) an amended Simberloff FORTRAN program, 2) a case study demonstrating the feasibility of this modified program, and 3) a

comparison of the validity of describing species diversity by employing rarefaction measurement and conventional indices (H' and J'). The studied communities are benthic polychaete communities located off the coast of southwestern Taiwan.

MATERIALS AND METHODS

Rationale of rarefaction

Rarefaction is a procedure for analyzing the number of species (species richness) among collections, when all collections are scaled down to the same number of individuals. This scaling procedure was termed 'rarefaction' by Sanders (1968) and was improved upon by Hurlbert (1971). The number of species, S_n , that can be expected from a random sample of n individuals, drawn without replacement from N individuals distributed among S species, is given by

$$E(S_n) = \sum_{i=1}^S \left[1 - \binom{N - N_i}{n} / \binom{N}{n} \right]$$

where S is the total number of species found in the collection, and N_i is the number of individuals of the i th species (Hurlbert 1971). The formula computes the expected number of species in a random sample of n individuals as the sum of the probabilities that each species will be included in the sample (James and Rathbun 1981).

The variance of $E(S_n)$ is given by Heck and coauthors (Heck et al. 1975) as

$$\text{Var}(S_n) =$$

$$\sum_{i=1}^S \left\{ \left[\binom{N - N_i}{n} / \binom{N}{n} \right] \left[1 - \binom{N - N_i}{n} / \binom{N}{n} \right] \right\} \\ + 2 \sum_{i < j}^S \left\{ \left[\binom{N - N_i - N_j}{n} / \binom{N}{n} \right] \right. \\ \left. - \left[\binom{N - N_i}{n} / \binom{N}{n} \right] \left[\binom{N - N_j}{n} / \binom{N}{n} \right] \right\}$$

calculated following an approximation estimation given by Pielou (1966b). The formulae of the Shannon index and the approximation of variance are given, respectively, as

$$H' = - \sum_{i=1}^s \left[\left(N_i / N \right) \ln \left(N_i / N \right) \right]$$

and

$$\text{Var} (H') \cong (1 / N) \left[\sum_{i=1}^s \left(N_i / N \right) \left(\ln \left(N_i / N \right) \right)^2 - (H')^2 \right].$$

Rarefaction measurements were generated using our FORTRAN IV program amended from the Simberloff program (Simberloff 1978). The differences in the numbers of species expected and the Shannon indices between the 2 regions were compared at a significance level of 5%.

RESULTS

Critique of the Simberloff and the Krebs programs

The Simberloff and the Krebs FORTRAN programs were written for mainframe computers (Simberloff 1978, Krebs 1989), such as a CDC Cyber 73 Computer, which has a large amount of memory capacity. Numerical calculation of numbers, such as

$$T(k) = \sum_{i=1}^k \log (i) \quad \text{for } k = 1, 2, \dots, 6000$$

requires large amounts of memory space for storage. Mathematically, $T(k)$ goes to infinity very quickly at the rate $k \log(k)$, as k approaches infinity. As a result, numerical overflow problems often occur, especially when using personal computers which have limited memory space.

Note that both the formula of the expectation and of the variance of the number of species S_n have terms like

$$\binom{N-H}{n} / \binom{N}{n}$$

where $H = N_i$ or $N_i + N_j$ for some i, j . This can be rewritten as

$$P(H) = \binom{N-H}{n} / \binom{N}{n}$$

$$\begin{aligned} &= \frac{(N-H)! / n! (N-H-n)!}{N! / n! (N-n)!} \\ &= \frac{(N-H)(N-H-1) \dots (N-H-n+1)}{N(N-1) \dots (N-n+1)} \\ &= \frac{N-H}{N} \times \frac{N-H-1}{N-1} \times \dots \times \frac{N-H-n+1}{N-n+1} \\ &= \prod_{j=1}^n \frac{N-H-j+1}{N-j+1} \end{aligned}$$

Hence, $\binom{N-H}{n} / \binom{N}{n}$ can be expressed as a product of n terms, where each term $\frac{N-H-j+1}{N-j+1}$ is no greater than 1 and no less than $\frac{N-H}{N}$, and bounded away from zero. Therefore, $E(S_n)$ and $\text{Var}(S_n)$ can be rewritten as follows

$$E(S_n) = \sum_{i=1}^s [1 - P(N_i)]$$

$$\begin{aligned} \text{and } \text{Var}(S_n) &= \sum_{i=1}^s [P(N_i) (1 - P(N_i))] \\ &\quad + 2 \sum_{i < j} [P(N_i + N_j) - P(N_i) \cdot P(N_j)]. \end{aligned}$$

In the above calculation, only S numbers of $P(N_i)$'s need to be stored, where S is the number of total species in the collection. It is obvious that the amount of computer memory utilized is far smaller than N , the total number of individuals, in the collection.

The amended rarefaction FORTRAN program is provided in the Appendix, and an instruction manual for the program is available (Li and Hsieh 1997) from the 2nd author upon request.

The case study: rarefaction measurements of polychaete communities

A total of 58 polychaete species were found in the study area, and 26 species were shared by both regions. In the north region, 724 individuals of 56 species were collected, and in the south region 378 individuals of 29 species were collected. The number of species occurring at each station varied from 0 to 18, with an average of 4.8 species at each station (Fig. 2).

The rarefaction curves of the 2 regions show that the number of species expected in the north

region is greater than that for the south region. In addition, the 95% confidence limits do not overlap with each other, except in samples having less than 25 individuals (Fig. 3), indicating that polychaete species diversity is richer in the north than in the south. The curve of the polychaete community is steeper for the north region than for the south, suggesting a more even distribution of abun-

dance among polychaete species in the north region than in the south (Fig. 3).

The case study: the Shannon and evenness indices

For the north region, the Shannon species diversity index (H') was 3.29, with the approximation of 95% confidence limits ranging from 3.20 to 3.37, whereas in the south region the diversity index was 2.10, with the approximation of 95% confidence limits ranging from 1.95 to 2.24. The evenness index was 0.82 in the north region and 0.52 in the south region. The confidence limits associated with the Shannon index of each region seem not to overlap, suggesting that diversity is greater in the north region than in the south.

DISCUSSION

Strength of the amended rarefaction program

As the rarefaction formula is re-expressed, the calculations of the intermediate term ($P(N_i)$), the expected number of species ($E(S_n)$), and its variance ($Var(S_n)$) do not involve special functions, such as logarithm (log) or raising e to a power (exp); rather, operations employed are restricted to addition, subtraction, multiplication, and division (see Appendix). In addition, the new program needs only 1 extra variable of dimension S in addition to S num-

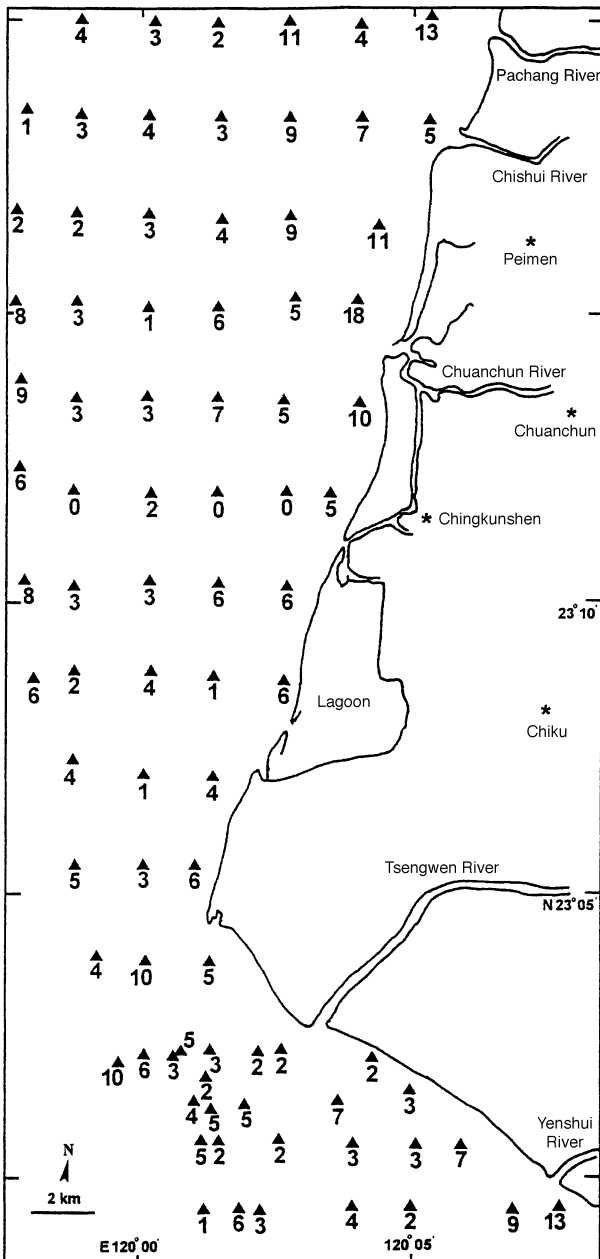


Fig. 2. Distribution of the number of species occurring at each station.

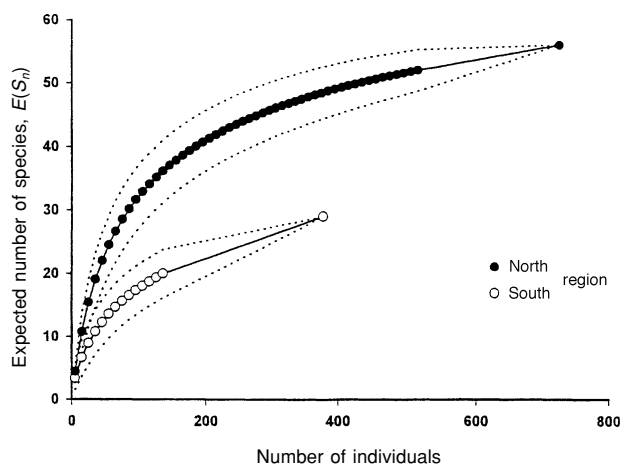


Fig. 3. Comparisons between the north region and the south region of species diversity in polychaete communities using rarefaction measurements. Dashed lines depict the 95% confidence limits ($\pm 1.96 Var^{1/2}$).

bers of $P(N)$. In contrast, the Simberloff and the Krebs programs (Simberloff 1978, Krebs 1989) need memory space to store not only 6000 $T(k)$'s but also 4 extra variables; each dimension by default is 550, or the total number of species, S . Hence, the new program uses less memory space and works more efficiently.

Comparisons of measurements in species richness and diversity

Species diversity is determined by the number of species (species richness), the relative abundance of those species (evenness), and the number of individuals sampled (sample size) (e.g., Pielou 1966b, Smith and Grassle 1977, James and Rathbun 1981). In the study area, the species richness is significantly greater in the north region than in the south region when the 2 collections differing in sample size are rarefied to the same number of individuals (see Fig. 3). The species evenness is also greater in the north region, as the rarefaction curve is steeper (see Fig. 3). Although the Shannon and evenness indices of the polychaete communities are greater in the north region than in the south ($H' = 3.29$ vs. 2.10 , $J' = 0.82$ vs. 0.52), the Shannon index has been regarded as inappropriate for interpreting species diversity (e.g., Pielou 1966b, Heck et al. 1975, Smith and Grassle 1977). The index is a biased estimation of the true population diversity and has been heavily criticized (e.g., Pielou 1966b, Hurlbert 1971). In addition, one could argue that large samples would contain more species than small ones (Heck et al. 1975), and a single index value confounds species richness, species evenness, and sampling size or area (James and Rathbun 1981). A single value does not show the whole range of changes in species richness and the relative distribution of abundance in a collection. Thus, the greater Shannon and evenness indices for the north region may not reflect the truth, due to the aforementioned pitfalls associated with the Shannon diversity formula. In contrast, rarefaction method analysis demonstrates that the polychaete species diversity is indeed greater in the north than in the south region.

The Shannon index method can be used if comparisons are based on equal sample sizes. It needs to be stressed that, in the present case study, 3 measurements, species richness, evenness, and sample size, are all greater in the north than in the south region, and the resulting Shannon diversities show the same conclusion as the rare-

faction diversity does. However, if 1 of these 3 measurements does not hold the same trend between the 2 regions, the conclusion varies with the method used. More importantly, the Shannon method can not differentiate which of the measurements makes the difference.

Applications of rarefaction method

Rarefaction has been used appropriately not only in describing community structure, but also in testing whether different samples have been drawn from the same community. Rarefaction also can be used to answer questions related to taxonomic diversity in evolutionary ecology. Several examples have been elucidated by Simberloff (1978).

The amended program presented in this study has advantages in its practical uses. After the program is converted to an executive file by a FORTRAN compiler, the program works on any personal computer where the FORTRAN language is no longer needed. The program can be installed in small computers, such as a notebook personal computer, and therefore, benefits field ecological research.

Some limitations of using rarefaction

Rarefaction examines the numbers of individuals, not the identities of species; different samples rarefied to the same number of individuals might have the same number of species, but have none of the same species (Simberloff 1978). Rarefaction compares the samples which should be collected with similar methods and from similar habitats and which belong to similar taxa (Krebs 1989). In addition, sessile organisms tend to be clumped; under-dispersion is likely to result in overestimating the number of species expected (Simberloff 1978). In the present case study, polychaetes are, in general, sessile organisms, but a relatively large area was sampled in each region (approximately 30×10 km² in the north region and 6×12 km² in the south), resulting in collections from a number of different clumps, and thus, the problem of spatial under-dispersion was avoided.

Acknowledgments: Hwey-Lian Hsieh wishes to thank K.S. Lee, P.J. Liu and S.T. Shi for their assistance in field and laboratory work, and Chang-Po Chen for his ongoing constructive comments. This study was supported by NSC Grant NSC842621P001006.

REFERENCES

- Grassle JF, NJ Maciolek. 1992. Deep-sea species richness: regional and local diversity estimates from quantitative bottom samples. *Am. Nat.* **139**: 313-341.
- Heck KL Jr., G van Belle, D Simberloff. 1975. Explicit calculation of the rarefaction diversity measurement and the determination of sufficient sample size. *Ecology* **56**: 1459-1461.
- Hurlbert SH. 1971. The nonconcept of species diversity: a critique and alternative parameters. *Ecology* **52**: 577-586.
- James FC, S Rathbun. 1981. Rarefaction, relative abundance, and diversity of avian communities. *The Auk* **98**: 785-800.
- Krebs CJ. 1989. *Ecological methodology*. New York: Harper Collins Publisher.
- Li LA, HL Hsieh. 1997. Programs of the rarefaction, technical report C-97-03. Taipei, Taiwan: Institute of Statistical Science, Academia Sinica.
- Ludwig JA, JF Reynolds. 1988. *Statistical ecology*. New York: J. Wiley.
- MacArthur RH. 1965. Patterns of species diversity. *Biol. Rev.* **40**: 510-533.
- Pianka ER. 1966. Latitudinal gradients in species diversity: a review of concepts. *Am. Nat.* **100**: 33-46.
- Pielou EC. 1966a. The measurement of diversity in different types of biological collections. *J. Theor. Biol.* **13**: 131-144.
- Pielou EC. 1966b. Shannon's formula as a measure of specific diversity: its use and misuse. *Am. Nat.* **100**: 463-465.
- Sanders HL. 1968. Marine benthic diversity: a comparative study. *Am. Nat.* **102**: 243-282.
- Simberloff D. 1978. Use of rarefaction and related methods in ecology. *In* KL Dickson, J Cairns Jr., RJ Livingston, eds. *Biological data in water pollution assessment: quantitative and statistical analyses*. ASTM STP 652. Easton, Md: American Society For Testing and Materials, pp. 150-165.
- Smith W, JF Grassle. 1977. Sampling properties of a family of diversity measures. *Biometrics* **33**: 283-292.
- Wilson EO. 1994. Biodiversity: challenge, science, opportunity. *Amer. Zool.* **34**: 5-11.

Appendix

```

PROGRAM RAREFACTION
C   *** This is the RAREFACTION PROGRAM to provide estimated means
C   (SM) and variances (V) of the number of species when you
C   randomly draw N individuals from the total individuals
C   of size NN.
C   Before you run this program, please specify the input and
C   output data sets in the 'OPEN' statements, and modify
C   the possible total number of species in the 'DIMENSION'
C   statement, where 600 is the default value in this program.
C
C   IMPLICIT REAL*8(A-H,O-Z)
C   DIMENSION TM(600),IN(600)
C   *** You might replace 600 by any possible total number of
C   species of your data.
C
OPEN(1,FILE='rf.dat',STATUS='OLD')
C   *** Here, "rf.dat" is the input data file containing one column
C   of the number of individuals in each species. Of course,
C   you can use name other than "rf.dat" as your input data
C   file. If you have more than one variable, please modify
C   the "READ" statement numbered 10 below.
C
OPEN(2,FILE='rf.out',STATUS='UNKNOWN')
C   *** Here, "rf.out" is the output data file containing
C   the number of species, the number of individuals,
C   sample size n, means and variances.
C
C   NN = the total number of individuals
C   NS = the total number of species in NN individual
C   N = sample size drawn randomly from NN individuals
C   IN(J) = the number of individuals in the Jth species
C
C   NS=0
C   NN=0
C   J=1
10  READ(1,*,END=11)IK
C   NS=NS+1
C   NN=NN+IK

```

```

IN(J)=IK
J=J+1
GO TO 10
C
11 WRITE(*,*)'There are ',NS,' Species'
WRITE(*,*)'There are ',NN,' Total Individuals'
WRITE(2,*)'There are ',NS,' Species'
WRITE(2,*)'There are ',NN,' Total Individuals'
WRITE(*,*)
C
C *** IA is the initial sample size, IB is the final sample size,
C and IC is the increment from IA to IB of sample sizes N in which
C you might be interested to have their means and variances.
C For example, if you would like to estimate means and variances
C for sample sizes N = 60,100,... 500. You could hit
C 60 <RETURN> 500 <ENTER> 40 <ENTER>
C following the request appearing on screen.
C
20 WRITE(*,*)'Please enter the smallest sample size,'
WRITE(*,*)' the largest sample size and increment of'
WRITE(*,*)'sample size n you consider to draw.'
READ(*,*)IA,IB,IC
WRITE(*,*)
WRITE(2,*)' N MEAN VARIANCE'
WRITE(*,*)' N MEAN VARIANCE'
DO 40 N=IA,IB,IC
SM=0
V=0
DO 30 K=1,NS
V1=0
TM(K)=1
DO 21 J=1, N
IF ((NN-IN(K)-J+1).LE.0) GO TO 50
TM(K)=TM(K)*(FLOAT(NN-IN(K)-J+1)/FLOAT(NN-J+1))
21 CONTINUE
DO 25 I=1,K-1
V2=1
DO 22 J=1, N
IF ((NN-IN(I)-IN(K)-J+1).LE.0) GO TO 50
V2=V2*(FLOAT(NN-IN(I)-IN(K)-J+1)/FLOAT(NN-J+1))
22 CONTINUE
V1=V1+V2-TM(I)*TM(K)
25 CONTINUE
SM=SM+1-TM(K)
V=V+TM(K)*(1-TM(K))+2*V1
30 CONTINUE
WRITE(2,3)N,SM,V
3 FORMAT(I5,2X,2F10.4)
WRITE(*,3)N,SM,V
40 CONTINUE
IF (N.GT.IB) GO TO 60
50 WRITE(*,*)
C The following message will appear on the screen when sample size n is
C so large that it makes either (NN-IN(K)-J+1) or (NN-IN(I)-IN(K)-J+1)
C equal to or smaller than 0. The output data file will not include
C means and variances for these large n's.
C
WRITE(*,*)N,' is too large. Please use n less than ',J
WRITE(2,*)
CONTINUE
C
60 WRITE(*,*)
WRITE(2,*)
WRITE(*,*)
WRITE(*,*)'Do you need another rarefaction estimation?'

```



```
WRITE(*,*)  
WRITE(*,*)' (1) Yes, I need another one.'  
WRITE(*,*)' (2) No. I want to quit now.'  
WRITE(*,*)  
WRITE(*,*)' Please enter your selection: 1 or 2,'  
WRITE(*,*)' then press <ENTER>'  
READ(*,*) MA  
IF (MA.EQ.1) GO TO 20  
CLOSE (1)  
CLOSE (2)  
STOP  
END
```

歧異度：修正之稀釋法 FORTRAN 程式應用於多毛類群聚

謝蕙蓮¹ 李隆安²

本研究以多毛類群聚為實例，分析稀釋歧異度的意義與優點。報告了修訂 Simberloff 氏所寫 FORTRAN 稀釋曲線程式，並用此新修程式比較多毛類群聚之種歧異度。經重新展開稀釋曲線數學式後，發現 Simberloff 氏之程式可簡化。簡化後之程式在計算期望種數與其變方時，所需電腦記憶空間大幅縮減，但效率卻大為提升。本論文的程式，經 FORTRAN 語言編譯器轉換成執行檔後，可以放在沒有 FORTRAN 語言的個人電腦上執行，對田野生態工作提供便利及助益。以棲息於臺南縣曾文溪沿岸亞潮帶地區之多毛類群聚為實例，比較南北兩區多毛類之種歧異度。採樣時間為 1994 年 9 月以及 1995 年 4 月。在北區總計採得 56 種多毛類，724 個個體；在南區，則計採得 29 種，378 個個體。當兩區個體數以稀釋法量化到相同個體數時，北區之期望種數多於南區。同時，北區的稀釋曲線斜率較南區的來得陡，顯示北區的多毛類個體數量在種間的相對分布較南區來得平均。另外，本研究也計算出傳統所用之種歧異度指數 (H') 與均勻度指數 (J')，以供稀釋曲線法比較之用。

關鍵詞：種歧異度，稀釋法電腦程式，底棲群聚。

¹中央研究院動物研究所

²中央研究院統計科學研究所