Original Version (November, 2003)
Latest updated (May, 2010)

User's Guide for Program SPADE (Species Prediction And Diversity Estimation)

by

Anne Chao, National Tsing Hua University, TAIWAN 30043
Tsung-Jen Shen, National Chung Hsing University, TAIWAN 402
Edited by Bruno A. Walther

Table of Contents:

Overview

The Program SPADE (Species Prediction And Diversity Estimation), written in C language, estimates various biodiversity indices based on different types of sample data from one or multiple communities. It is a free software, and this user guide attempts to explain how to use this program in an easily accessible way using numerical examples and explanations. The program is divided into seven parts:

- Part I: Species (estimating species richness for one community based on the observed species frequencies or presence/absence data)
- Part II: Shared Species (estimating the number of shared species for two communities based on frequency data or multiple incidence data)
- Part III: Prediction (predicting the number of new species that would be discovered in a second survey, based on frequency data from an initial survey)
- Part IV: Diversity Index (estimating various diversity indices including species richness, Fisher's alpha index, Shannon's entropy and Simpson's index as well as their effective numbers of species based on sample abundance or frequency data)
- Part V: Two-Community Similarity Index (estimating various similarity indices for two assemblages based on abundance data or multiple incidence data. The incidence-based indices include the classic Jaccard, Sørensen and Lennon et al. (2001) indices; the abundance-based indices include the Bray-Curtis, Morisita-Horn and two newly developed abundance-based Jaccard and Sørensen indices)

- Part VI: Multiple-Community Diversity Measure (estimating a class of generalized Morisita similarity/dissimilarity indices for more than two communities)

- Part VII: Genetics (applying Part VI to estimate allele similarity and differentiation for multiple-subpopulation genetic data)

The running procedures for each part are illustrated with numerical examples. There are several estimators and predictors available in each part. The formulas for each estimator or predictor with relevant references are provided in the Appendix. The user is referred to Magurran (1988, 2004) and Chao (2005) for background information.

Please do not use SPADE in any commercial form or distribute it to other people but direct them to the SPADE website (see below). If you publish your work based on results from SPADE, please make references to our relevant papers mentioned in the following sections and also use the following reference for citing SPADE:

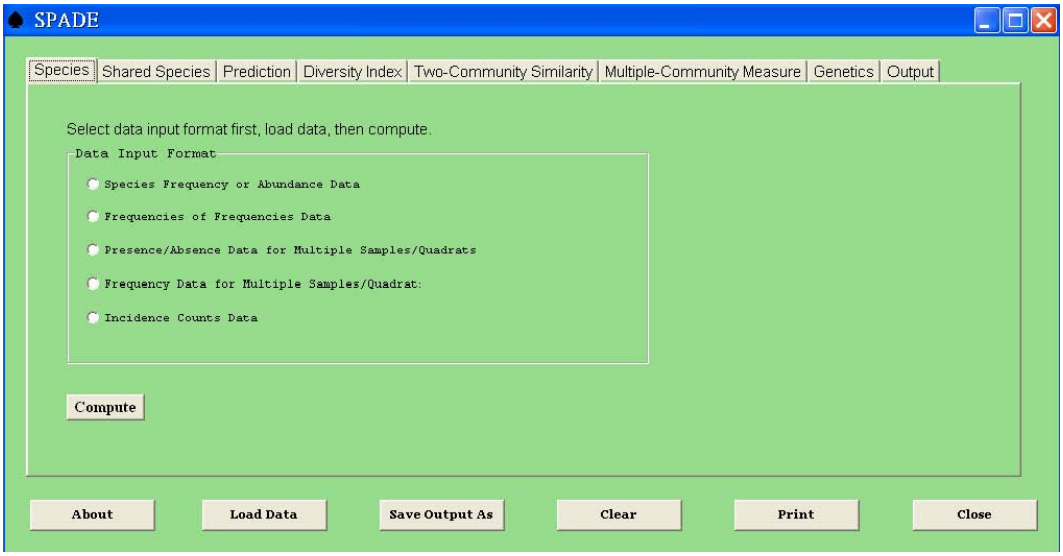
Chao, A. and Shen, T.-J. (2010) Program SPADE (Species Prediction And Diversity Estimation). Program and User's Guide published at http://chao.stat.nthu.edu.tw. (You can download the pdf files of all relevant papers directly from the above website.)

Download and Setup

The program SPADE can be downloaded free of charge from Anne Chao's website at http://chao.stat.nthu.edu.tw/softwareCE.html. In order to keep a record and contact you for future updated versions/information, you are asked to register before downloading. First download the program "SpadeInstall.exe", then double-click the same executable file to install the program. The source files along with several illustrative example data sets (in a sub-directory named "data") will be stored automatically in the default directory (C:\Program Files\Spade) on your computer. You can change the default directory to any other designated one. To gain familiarity with the program, we suggest that you first run the example data sets and check the output with that given in this guide. Part of the output for each example is also discussed in this user's guide to help the user to interpret the numerical results.

After the setup, double-click the executable file "SPADE.exe" to start the program with the interface window shown below in Figure 1. Along the top, there are eight selection tabs: Species (Part I), Shared Species (Part II), Prediction (Part III), Diversity Index (Part IV), Two-Community Similarity (Part V), Multiple-Community Measure (Part VI), Genetics (Part VII) and Output.

Figure 1. The Interface Window of SPADE



Data Input Formats

All data files must be saved as ascii text files because SPADE can only handle tab-delimited or space-delimited ascii text files, except for data files which represent species data in rows, which also need a "return" (¶) at the end of each row. Therefore, if your data is stored in a database or spreadsheet program, save your data file as an ascii text file. In other words, data entries (meaning the actual numbers) must be separated by at least one blank space or a tab. If you have comma-delimited data (or data delimited with other symbols, e.g., semicolons), you need to replace them with the symbol for "blank space" ( ) or "tab" (→) in a word processing program using the find-and-replace function.

The following sections explain how different ascii text files need to be formatted so that SPADE can read them properly. Therefore, the following sections concern only how the data files must be structured; their use when actually running the analyses will be explained in later sections (beginning with "Running Procedures by Examples").

These sections on data format are divided into three parts: the first part explains data formats for "One Community/Assemblage" data, the second part explains data formats for "Two Communities/Assemblages" data, and the third part explains data formats for "More Than Two Communities/Assemblages" data.

Not every one of these three different data formats can be used for every one of the available analyses of the SPADE program. These different analyses are subdivided into seven parts, called Species, Shared Species, Prediction, Diversity Index, Two-Community Similarity, Multiple-Community Measure, and Genetics (corresponding to the seven tabs at the top of the interface). Below is a table specifying which data format can be used for which analyses:

There are five types of data formats:
Type (1): Species Frequency or Abundance Data:
Type (2): Frequencies of Frequencies (Frequency Counts) Data
Type (3): Presence/Absence Data for Multiple Samples/Quadrats
(i.e., Multiple Incidence Data)
Type (4): Frequency Data for Multiple Samples/Quadrats

Type (5): Incidence Counts Data for Multiple Samples/Quadrats

Analyses part	**Data formats options**
Part I: Species	One community, Type (1), (2), (3), (4), and (5)
Part II: Shared Species	Two communities, Type (1) and (3)
Part III: Prediction	One community, Type (1) and (2)
Part IV: Diversity Index	One community, Type (1) and (2)
Part V: Two-Community Similarity	Two communities, Type (1) and (3)
Part VI: Multiple-Community Measure	Multiple communities, Type (1)
Part VII: Genetics	Multiple communities, Type (1)

One Community/Assemblage:

When your data originate from only one community or assemblage, the program accepts five different types of data input formats:

Type (1): Species Frequency or Abundance Data:

For any species found in the sample, this type of data includes the number of times (frequency) that the species was discovered, or the number of individuals (abundance) that the species was represented in sample. We use a set of birds data from Magurran (1988, p. 152) as an illustrative example; the data are stored in **Data1a.txt** in the accompanied data files (in the default folder C:\Program Files\SPADE\Data). These data files can be viewed in any word processing program (e.g., Notepad, Microsoft WORD). The data set contains 25 observed species, and their observed frequencies are respectively 752, 276, 194, 126, 121, 97, 95, 83, 72, 44, 39, 16, 15, 13, 9, 9, 9, 8, 7, 4, 2, 2, 1, 1, 1 (that is, the first species is represented by 752 individuals, the second species by 276 individuals, ..., and the last three species are each represented by only one individual). The required data entry format is shown below. First, the frequency data must be entered in a *column*. Second, if species names are recorded in your original data, they must be removed to conform to the data format shown below. Third, rows with *frequency 0 for an unobserved species* may be included, but that species will not have any effect in the analysis. We specifically add some zero's in the preceding frequencies in **data1a.txt** for illustration. Frequencies do not have to be ordered according to magnitude, as can be seen in the example below. For data where each row denotes the number of individuals of a species (i.e., the species' abundance or frequency), the number needs to be followed by a return symbol, as in the example below.

752
276
194
126
121
97
95
83
72
44
39

0
16
15
0
13
9
9
9
8
7
4
0
0
2
2
1
1
1

Type (2): Frequencies of Frequencies (Frequency Counts) Data

Species frequency data are often classified by their frequencies into a simple form of $(f_1, f_2,..., f_r)$, where r denotes the maximum frequency and f_k denotes the number of species represented by exactly k individuals/times in the sample and r denotes the maximum frequency. The statistics $(f_1, f_2,..., f_r)$ are often referred to as "frequencies of frequencies" or "frequency counts". As an example, for the above bird data set, $f_1 = 3$ (there are three species represented by one individual), $f_2 = 2$ (there are two species represented by two individuals), $f_3 = 0$ (no species represented by three individuals), $f_4 = 1, f_5 = 0, ..., f_{752} = 1$, with the maximum frequency $r = 752$. Again, for data entry, we only require the non-zero values of f_k's, but you may include zero values; however, then you need to increase the number of pairs m accordingly. Data must be arranged in the following order: $(r, m, 1, f_1, 2, f_2,..., r, f_r)$, where r denotes the maximum frequency and m denotes the number of values of f_k's (i.e., the number of frequencies). The data entry for the bird data from above is $(r, m, 1, f_1, 2, f_2,..., r, f_r)$ = (752 20 1 3 2 2 4 1 7 1... 752 1); here, each number needs to be separated by at least one blank space. Here m also equals the number of pairs that follow the number m in the data sequence.

The frequencies of frequencies for coin data stored in **Data1b.txt** are taken from Holst (1981) and Chao and Lee (1992): 204 coins were found in a hoard of ancient coins. They were classified to different die types, with dies being an engraved metal piece used for impressing the coin's design. Here, the goal was to estimate the number of dies in the minting process. Here a "species" means a "different die type". The frequency counts for the obverse sides (disregarding the other side) are: $f_1 = 102, f_2 = 26, f_3 = 8, f_4 = 2, f_5 = f_6 = f_7 = 1$. That is, 102 dies appeared on only one coin, 26 appeared on two coins, and so on, with 1 die appearing on 7 coins (each coin had only one die) which is therefore the maximum frequency r. The data $(r, m, 1, f_1, 2, f_2,..., r, f_r)$ are read as:

7 7 1 102 2 26 3 8 4 2 5 1 6 1 7 1

whereby m is again the number of data pairs following m, which do not have to appear in any particular order (even though the example above is ordered). This particular example is found

in the data file **Data1b.txt**; the seven pairs are (1, 102), (2, 26), (3, 8), (4, 2), (5, 1), (6, 1) and (7, 1).

Type (3): Presence/Absence Data for Multiple Samples/Quadrats (i.e., Multiple Incidence Data)

In some ecological studies, incidence (presence/absence) data are collected over repeated samples in time and/or space. Quadrat sampling provides an example in which the study area is divided into a number of quadrats, and a sample of quadrats is randomly selected for observation. There are other similar examples: sampling is conducted by several investigators, or trapping records are collected over multiple occasions. From hereon, we use the general term "sample/quadrat" or simply "sample" to refer to the various possible sampling units, e.g., quadrats, occasions, sites, transect lines, periods of fixed time, fixed number of traps, investigators, and so on. Therefore, SPADE can also be used for estimating the population size based on temporally replicated samples (e.g., capture-recapture studies) if marking or tagging enables the researcher to distinguish individuals. In this case, each individual is treated as a species, and the species frequency then corresponds to the number of times that an individual was sighted or captured.

Thus, for this type of data, a number of samples are collected. In each sample, only the presence/absence of each species is recorded. Here, we use the seed-bank data analyzed in Colwell and Coddington (1994) for illustration which contains 121 standardized soil samples collected from one-hundred 10 m x 10 m grids in a Costa Rican forest. A total of 34 species of seedlings germinated from these 121 samples. As shown in the data file **Data1c.txt**, the data are arranged in a matrix with 34 rows and 121 columns. The presence of any species in a sample is denoted by 1 and its absence is denoted by 0. The data were kindly provided by Dr. Colwell.

Type (4) : Frequency Data for Multiple Samples/Quadrats

The format is similar to the matrix for presence/absence data under (3) except that each data entry in the data matrix denotes the sample abundance or frequency instead of incidence. For example, in the seedlings data, the number of individuals for each observed species were also provided by Dr. Colwell for each sample. These abundance counts can be found in **Data1d.txt**. The data entry format is again a 34 x 121 matrix, but each cell denotes species frequency or abundance in the sample.

Type (5): Incidence Counts Data for Multiple Samples/Quadrats

The presence/absence data for t samples/quadrats are often summarized by incidence counts $(Q_1, Q_2,...,Q_r)$, where Q_k denotes the number of species that were detected in exactly k samples/quadrats in the data; and r denotes the number of samples/quadrats in which the most frequent species are found. Thus, Q_1 and Q_2 represent the number of species that occur in exactly one sample ("uniques") or in exactly two samples ("duplicates"), respectively. In data entry, we require only the non-zero value of Q_k's. Data are arranged in the following order: $(t, r, m, 1, Q_1, 2, Q_2,..., r, Q_r)$ where m denotes the number of values of Q_k's; however, you may include zero values, but then you need to increase the number of pairs m accordingly. For example, the incidence counts data $(t, r, m, 1, Q_1, 2, Q_2,..., r, Q_r)$ of the seedbank data are read as:

121 61 18 1 3 2 2 3 3 4 3 5 1 6 5 7 1 8 1 9 3 10 1
11 2 13 1 17 1 24 2 43 2 47 1 52 1 61 1

(See **Data1e.txt).** The first entry 121 means there are 121 soil samples; the second entry means that the most frequent species is found in 61 samples; and the third entry means that there are 18 incidence counts. The third entry 18 also implies that the remainder of the dataset consists of 18 number pairs which are (1, 3), (2, 2), (3, 3), (4, 3), (5, 1), (6, 5), and so on up to (61, 1). Thus (1, 3) indicates that there are 3 unique species, (2, 2) indicates there are 2 duplicate species, and so on, and (61, 1) indicates that there is one species found in 61 soil samples.

Two Communities/Assemblages:

When your data originate from two communities or assemblages, the program accepts two different types of data input formats:

Type (1): Species Frequency or Abundance Data

Assume that in two communities, one (and only one) sample of individuals is taken, each sample containing different numbers of individuals for several species. These data must be stored as a species (in rows) by community (in two columns) matrix file. The first column denotes the frequency (or abundance) of a species discovered in Community I, and the second column denotes the frequency (or abundance) of the same species in Community II. The two frequencies are separated by at least one blank space or a tab. For example, the data stored in **Data2a.txt** are bird frequency records of two estuaries. The data are arranged as follows:

39	42
0	70
0	1
842	616
.	.
.	.
.	.
1	1

In this data set, the first species was observed 39 times in Community I and 42 times in Community II, the second species was not observed in Community I whereas it was observed 70 times in Community II, and the last species was observed once in each community. Each row lists the frequencies or abundances for a specific species and different rows refer to frequencies of different species. SPADE cannot handle missing data, which are empty cells without numbers. Therefore, when a species was not observed in a community, you must enter the frequency 0 (as, e.g., in the second and third row of the first column above). Again, SPADE cannot handle species names so they need to be removed from the data file.

Type (3): Multiple Incidence Data

Assume that in two communities, several independent samples are recorded of the presence/absence of a number of species. Therefore, in contrast to the example above (1), here we enter the total number of presences for each species (which cannot exceed the total number of independent samples). For example, we here use the Hong Kong Big Bird Race (BBR) Data which are stored in the file **data2b.txt**. The Hong Kong Big Bird Race (BBR) is an annual competition among teams of birdwatchers. The challenge is to record as many bird species as possible during a fixed interval of time in the Hong Kong territory. For example, 19 teams competed in 1999 while 20 teams competed in 2000 (see **data2b.txt**):

19	20
0	0
0	0

.	.
1	0
.	.
.	.
19	20
19	20

In this data format, the first row denotes the number of samples (19 and 20 teams, respectively). Beginning with the second row, each row denotes the total number of presences of each species in all the samples. To illustrate, in the above data, the first and second species in the bird list were not observed by any team in 1999 and 2000. Meanwhile, a species entered as (1, 0) was observed by only one team in 1999 and was not observed by any team in 2000. Finally, the species in the bottom two rows were observed by all 19 and 20 teams in 1999 and 2000, respectively.

More Than Two Communities/Assemblages:

When your data originate from more than two communities or assemblages, the program accepts two different types of data input formats:

Type (1): Species Frequency or Abundance Data

This data format is simply an extended type of that for two communities (see above). Instead of two columns, we have *k* columns if there are *k* communities or assemblages. **For example,** the data stored in **Data5a.txt** are frequencies of an assemblage of seedlings (column 1), an assemble of saplings (column 2) and an assemblage of trees (column 3) recorded in the site called **LEP** old-growth rain forest in Costa-Rica (which is one of the sampling sites discussed in Chao et al. 2005, 2008; see these papers for more details on data).

17	48	7
14	38	0
.	.	.
.	.	.
1	0	0

In each of these *k* communities, one (and only one) sample of individuals is taken, each sample containing different numbers of individuals for several species. Again, these data must be stored as a species (in rows) by community (in columns) matrix file. The first column denotes the frequency (or abundance) of a species discovered in Community I, the second column denotes the frequency (or abundance) of the same species in Community II, and so on. The columns need to be separated by at least one blank space or a tab. In this data set, the first species is represented by 17 seedlings, 48 saplings and 7 trees; the second species is represented by 14 seedlings, 38 saplings and no trees; and the last species is represented by 1 seedling, no saplings and no trees. Again, missing data are not allowed, therefore you must store 0 in the data file if no observation was made. Again, currently SPADE cannot handle species names so they need to be removed from the data file. [In the analyses described in part V "Two-Community Similarity", the program offers you the option of choosing any combination of two communities (or assemblages) of the entire set of entered communities for similarity comparison.]

Type (3): Multiple Incidence Data

This data format is also simply an extended type of that for two communities (see above). Instead of two columns, we have *k* columns if there are *k* communities or assemblages. Thus in each of these *k* communities or assemblages, several independent samples are recorded of

the presence/absence of a number of species. Here, we use a data set of tropical rainforest ants collected in Costa-Rica (Longino et al. 2002) for illustrating the data format (**Data5b.txt**). Ants were captured by using three techniques: (a) Berlese extraction of soil samples (217 samples) (b) fogging samples from canopy fogging (459 samples), and (c) Malaise trap samples for flying and crawling insects (62 samples). Therefore, we here denote each sampling technique to represent a different assemblage, which are presented in three columns:

217	459	62
1	0	0
0	27	3
.	.	.
.	.	.
0	12	0

Again, in this data format, the first row denotes the number of samples (217, 459 and 62 denote the number of samples for Berlese, fogging and Malaise, respectively). **Beginning with the second row, each row denotes the total number of presences of each species in all the samples. To illustrate,** the first species was captured only in one Berlese sample and in none of the fogging and Malaise samples; the second species was not recorded in any of the Berlese samples, but in 27 fogging samples and 3 Malaise samples; and the last species was only captured in 12 fogging samples. [Again, in the analyses described in part V "Two-Community Similarity", the program offers you the option of choosing any combination of two communities (or assemblages) of the entire set of entered communities for similarity comparison.]

Genetic Data

Type (1): Frequency or Abundance Data

When your data are allele frequency data, the program accepts only one input format. We use allele frequency data as an example. Store your data as an allele (row) by subpopulation (column) matrix file. In many genetic data, only allele proportions are given. To use SPADE, you must convert proportional data to frequency data (the actual number of sampled instances of allele *j*). As an example, the file **Data7b.txt** includes allele frequency counts in locus D3S2427 from four human subpopulations (BiakaPyg, Palestin, Bedouin and Druze; data for further subpopulations are presented in reference, but for simplicity, we here only present data for four of those subpopulations):

0	6	3	1
11	0	0	0
5	0	0	0
.	.	.	.
.	.	.	.
2	0	0	0
0	0	0	0

For any allele which is not found in the sample from a subpopulation, you must enter the frequency 0 in the data file. For example, the last allele was not found in any of the four chosen subpopulations (although it was found in another subpopulation whose data are not shown in this example); therefore, we have four 0's in the last row.

Running Procedures by Examples

Part 1: Species (Species Richness Estimation in One Community)

In the following, we present running procedures using separate examples for each of the five types of data.

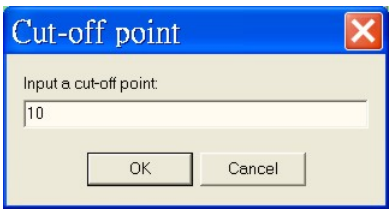
Example 1a (Species Frequency or Abundance Data) Birds Data in Data1a.txt

(The detailed description for this data set is given above in Data Input Formats (1) for one community.)

Step 1. Select proper data type from Data Input Formats (see Data Input Formats above, (1) Species Frequency or Abundance Data in this particular case).

Step 2. Load your data by pressing the key Load Data, then input the complete path of the data file by, for example, sample data sets can be found in the default folder C:\Program Files\SPADE\Data.

Step 3. Press the key Compute to calculate the estimates. The following window pops up:



The cut-off point is a value that separates frequency counts into abundant and rare groups. See the example below for an explanation of this cut-off value. The default value is 10 (which from our experience is an appropriate value for most data sets; however, a higher cut-off is recommended for microbial communities; see guidelines below). Press OK, then the window screen is minimized to

The computation is complete when the following window pops up:

Step 4: Press OK, then the output is shown on the screen under the "Output" tab. You can press the key Save Output As to save your output as a textfile or press Print to print a hard copy of your output.

OUTPUT:

```
(1)  BASIC DATA INFORMATION:

        (Number of observed individuals)          n =  1996
        (Number of observed species)              D =    25
        (Cut-off point)                           k =    10

     "Rare" Species Group: (Frequency counts up to the cut-off point)
          ------------------------------------------------------------------------
```

```
 f[1]=3; f[2]=2; f[4]=1; f[7]=1; f[8]=1; f[9]=3;
-------------------------------------------------------------------------
(Number of observed individuals for rare species)     n_rare   =    53
(Number of observed species for rare species)         D_rare   =    11
(Estimation of the sample coverage for rare species)  C_rare   = 0.943
(Estimation of CV for rare species in ACE)            CV_rare  = 0.629
(Estimation of CV1 for rare species in ACE-1)         CV1_rare = 0.740

"Abundant" Species Group: (Frequencies beyond the cut-off point)
(Number of observed individuals for abundant species)     n_abun =  1943
(Number of observed species for abundant species)         D_abun =    14
```

(2) ESTIMATION OF SPECIES RICHNESS:

Estimator/Model	Estimate	Est_s.e.	95% Confidence interval
Homogeneous Model	25.7	1.0	(25.1 , 30.3)
Homogeneous (MLE)	25.0	0.0	(25.0 , 25.0)
Chao1 (Chao, 1984)	27.3	3.4	(25.3 , 44.0)
Chao1-bc	26.0	1.8	(25.1 , 35.7)
ACE (Chao & Lee, 1992)	26.9	2.4	(25.3 , 37.6)
ACE-1 (Chao & Lee, 1992)	27.4	3.2	(25.3 , 42.2)
1st order jackknife	28.0	2.4	(25.7 , 37.2)
2nd order jackknife	29.0	4.2	(25.7 , 46.9)
Gamma-Poisson Model (Chao & Bunge, 2002)	27.5	4.5	(25.2 , 51.6)
Gamma-Poisson-UMLE	24.6	1.9	(25.0 , 11.9)
Gamma-Poisson-CMLE	*	*	(* , *)

* iterative steps do not converge.

(3) DESCRIPTION OF ESTIMATORS/MODELS:

Homogeneous Model: This model assumes that all species have the same detection probabilities.
See Eq.(2.3) in Chao and Lee (1992) or Eq. (2.1) in Chao et al. (2000).

Homogeneous (MLE): An approximate maximum likelihood estimate under homogeneous model.
See Eq.(1.1) and Eq.(1.2) in Chao and Lee (1992).

Chao1 (Chao, 1984): This approach uses the numbers of singletons and doubletons to estimate the number of missing species because undetected species information is mostly concentrated on those low frequency counts; see Chao (1984), Shen, Chao and Lin (2003) and Chao, Shen and Hwang (2006).

Chao1-bc: a bias-corrected form for the Chao1; see Chao (2005).

ACE (Abundance-based Coverage Estimator): A non-parametric estimator proposed by Chao and Lee (1992) and Chao, Ma and Yang (1993). The observed species are separated as rare and abundant groups; only the rare group is used to estimate the number of missing species.
The estimated CV is used to characterize the degree of heterogeneity among species detection probabilities. See Eq.(2.14) in Chao and Lee (1992) or Eq.(2.2) in Chao et al. (2000).

ACE-1: A modified ACE for highly-heterogeneous communities. See Eq. (2.15) in Chao and Lee (1992).

1st order jackknife: It uses the number of singletons to estimate the number of missing species; see Burnham and Overton (1978).

```
2nd order jackknife: It uses the numbers of singletons and doubletons to estimate the
    number of missing species; see Burnham and Overton (1978).

Gamma-Poisson Model: An estimator proposed by Chao and Bunge (2002) for the gamma-Poisson model
    in which species are detected in the sample according to a Poisson process and rates
    of the processes follow a gamma distribution.  This model is suitable for situations when the
    community is sampled for a fixed time interval.

Gamma-Poisson-UMLE: Unconditional MLE for the Gamma-Poisson model; see Chao and Bunge (2002).

Gamma-Poisson-CMLE: Conditional MLE for the Gamma-Poisson model; see Chao and Bunge (2002).

95% Confidence interval: A log-transformation is used so that the lower bound of the
    resulting interval is at least the number of observed species.  See Chao (1987).
```

The output shown above is divided into three parts. The **first part** shows the basic data information including the total number of individuals, the observed number of species and the selected cut-off point. If the default cut-off of 10 is used, then the observed species are separated into two groups: "rare" and "abundant"; the former group includes species observed at most 10 times and the latter includes those observed at least 11 times. Only the frequency statistics $(f_1,\ f_2,...,f_{10})$ of the rare species (as defined by the cut-off point, in this case 10) are used to estimate the number of undetected species. We do not use the abundant species for estimation because they would be observed in almost every sample regardless of the observer and thus carry negligible information about the undetected species. On the other hand, we here assume that the information contained in the already detected species carries useful information about the undetected species which we use to estimate these undetected species. The frequency counts $(f_1,\ f_2,...,f_{10})$ are displayed, and basic information for the "rare" and "abundant" group are displayed under the headings "Rare Species Group" and "Abundant Species Group".

The **second part** lists estimates, standard errors and confidence intervals for each model or estimator of species richness (we use the terms "model" and "estimator" for different procedures of species richness estimation, but essentially, both types return species richness estimates). For the moment, the program features eleven models/estimators; however, future versions may expand this list. Brief details of each model/estimator are provided below this list in the **third part**, but for more details, please refer to Bunge and Fitzpatrick (1993), Colwell and Coddington (1994) and Chao (2005). The formulas for each model/estimator are provided in the Appendix of this user's guide.

The following part up to the section head "Example 1b (Frequencies of Frequencies Type) Coin Data in Data1b.txt" is a brief discussion of some of these eleven models/estimators, which reflects our experience with those estimators we developed. Therefore, it is not an exhaustive account of all eleven models/estimators, but focuses on those non-parametric estimators that we developed. You could skip ahead if you just wanted to follow instructions for the program's use.

The estimator proposed in Chao (1984) is referred to as the **Chao1** estimator in the ecological literature, e.g., in the program EstimateS (Colwell, 1997) and in Colwell and Coddington (1994). See Chao (2005) for a bias-corrected version. The estimator Chao1 is a simple and useful estimator; it uses only the numbers of singletons and doubletons to estimate the number of undetected species (see Remarks 1-3 below). It was derived as a lower bound of species richness

in Chao (1984). However, Shen, Chao and Lin (2003) and Chao et al. (2006b) verified that for many datasets it is relatively good point estimator of the total species richness of a community, thereby justifying its use as a species richness estimator (and not just an estimator of the lower bound of species richness).

While the Chao1 estimator only uses the information contained in singletons (f_1) and doubletons (f_2), the estimator referred to as Abundance-based Coverage (**ACE**) estimator, proposed by Chao and Lee (1992) and modified by Chao, Ma and Yang (1993), uses also the information contained in the higher frequencies (f_3 up to $f_{\text{cut-off}}$); see Appendix or Eq. (2.14) in Chao and Lee (1992) for its formula. The **ACE-1** estimator is a modified ACE estimator for a highly heterogeneous community; see Appendix or Eq. (2.15) in Chao and Lee (1992) for its formula. For highly heterogeneous communities, we suggest to use the Chao1 estimator as a lower bound and ACE or ACE-1 as the point estimate of species richness (while for communities with low to moderate heterogeneity, Chao1 can also be used as point estimator).

A measure used to characterize the degree of heterogeneity among species detection probabilities for rare species is the coefficient of variation (CV) for the rare species group. A CV = 0 would mean that all rare species are homogeneous (i.e., they all have equal detection probabilities in the community). Therefore, the larger the CV, the greater the degree of heterogeneity for species detection probabilities in the rare species group. The CV plays an important role in the ACE approach; see Chao et al. (2000) and Appendix for the formulas. The estimator of the CV (which is always ≥ 0) is also provided in the **first part** of the output, see CV_rare for ACE and CV1_rare1 for ACE-1. For this example, based on ACE, the estimated CV = 0.629 (which is the numerical value for the estimator CV_rare) signifies moderate heterogeneity in species detection probabilities. Based on ACE-1, the estimated CV = 0.740 (which is the numerical value for the estimator CV1_rare) signifies slightly higher heterogeneity. We recommend to use the ACE-1 only if the CV_rare exceeds 0.8; otherwise, the ACE should be used. Therefore, the estimate of species richness calculated from the "Homogeneous Model" will underestimate the true number of species. The Chao1, bias-corrected Chao1 and ACE produce slightly higher estimates. Because of the value of CV_rare, we recommend to use the species richness of the ACE; consequently, we conclude that two species remain undetected while the 95% confidence interval is (25.3, 37.6). Rounding off, we obtain a 95% confidence interval ranging between 25 and 38 species. The maximum likelihood estimators Gamma-Poisson-UMLE and Gamma-Poisson-CMLE could not be used in this particular case because of the divergence of numerical iterations.

In most cases when using a default cut-off value of 10, SPADE will return species richness estimates with the following ordering: ACE-1 > ACE > Chao1. We suggest/recommend that you use a cut-off value >10 if (a) the ACE estimator value is lower than the Chao1 estimate, or (b) for extremely heterogeneous communities (e.g., bacterial communities), meaning any community with very unequal detection probabilities.

Remark 1: The Chao1 estimator for $f_2 > 0$ has the form of $\hat{S}_{Chao1} = D + f_1^2/(2f_2)$, where D denotes the number of distinct species observed in the sample. It is s a very accurate lower bound for any type of abundance distribution. In microbial communities, however, when the sample size is small relative to the total species richness, species richness estimators based on fitting a parametric model are generally not stable. Nevertheless, this lower bound can still be accurately estimated with the Chao1 estimator. In a highly heterogeneous community, it is thus generally more useful to provide an accurate lower bound than an unstable point

estimate of total species richness. The variance of this estimator is $\hat{\text{var}}(\hat{S}_{Chao1}) = f_2\left[\frac{1}{2}\left(\frac{f_1}{f_2}\right)^2 + \left(\frac{f_1}{f_2}\right)^3 + \frac{1}{4}\left(\frac{f_1}{f_2}\right)^4\right]$, which can be subsequently used for constructing a confidence interval. In the special case of $f_2 = 0$, the Chao1 estimator is modified to $D + f_1(f_1 - 1)/2$; see Remark 2.

Remark 2: If all species detection probabilities are assumed to be equal, the Chao1 estimator is a statistically valid point estimate of species richness, and its bias can be evaluated. A large portion of bias associated with the Chao1 lower bound comes from small values of f_2. In the special case of *equal species detection probabilities*, an unbiased or "bias-corrected" estimator of the Chao1 estimator has the following form

$\hat{S}^*_{Chao1} = D + f_1(f_1 - 1)/[2(f_2 + 1)]$ which we call **Chao1-bc**. Unlike the Chao1 estimator, it is always obtainable even in the case of $f_2 = 0$. The variance formula for this bias-corrected estimator **Chao1-bc** is valid only for $f_2 > 0$

$$\hat{\text{var}}(\hat{S}^*_{Chao1}) = \frac{f_1(f_1-1)}{2(f_2+1)} + \frac{f_1(2f_1-1)^2}{4(f_2+1)^2} + \frac{(f_1)^2 f_2(f_1-1)^2}{4(f_2+1)^4}.$$

In the special case of $f_2 = 0$, the formula of Chao1-bc is reduced to $\hat{S}^*_{Chao1} = D + f_1(f_1 - 1)/2$.

The variance for this special case is slightly different from the one above; the third term in the above variance formula vanishes, but an additional term should be added as follows (Chao et al. 2006):

$$\hat{\text{var}}(\hat{S}^*_{Chao1}) = \frac{f_1(f_1-1)}{2} + \frac{f_1(2f_1-1)^2}{4} - \frac{f_1^4}{4\hat{S}^*_{chao1}}.$$

Remark 3: So why should one not always use the "bias-corrected" form? Note in Remark 2 that *the bias-corrected estimator is nearly "unbiased" only for the model of equal species detection probabilities*. Under a heterogeneous case, on the other hand, it is impossible to evaluate the bias of the Chao1 estimator because the exact heterogeneity pattern is generally unknown. Only under a homogeneous case or very low heterogeneity, we should use the unbiased form as a legitimate point estimate of total species richness.

Example 1b (Frequencies of Frequencies Type) Coin Data in Data1b.txt

The detailed description for this data set is given in Data Input Formats (2) for one community. All the steps are the same as those for Example 1a. The output for the coin data is shown below.

OUTPUT:

```
(1)  BASIC DATA INFORMATION:

        (Number of observed individuals)          n =   204
        (Number of observed species)              D =   141
        (Cut-off point)                           k =    10

    "Rare" Species Group: (Frequency counts up to the cut-off point)
```

```
------------------------------------------------------------------------
 f[1]=102; f[2]=26; f[3]=8; f[4]=2; f[5]=1; f[6]=1; f[7]=1;

------------------------------------------------------------------------
(Number of observed individuals for rare species)     n_rare   =   204
(Number of observed species for rare species)         D_rare   =   141
(Estimation of the sample coverage for rare species)  C_rare   = 0.500
(Estimation of CV for rare species in ACE)            CV_rare  = 0.686
(Estimation of CV1 for rare species in ACE-1)         CV1_rare = 0.986

"Abundant" Species Group: (Frequencies beyond the cut-off point)
      (Number of observed individuals for abundant species)     n_abun =     0
      (Number of observed species for abundant species)         D_abun =     0

(2)  ESTIMATION OF SPECIES RICHNESS:

     Estimator/Model            Estimate    Est_s.e.  95% Confidence interval
    ------------------------------------------------------------------------
     Homogeneous Model             282.0        32.9         (230.8 ,362.5)
     Homogeneous (MLE)             258.1        25.0         (218.4 ,318.0)
     Chao1 (Chao, 1984)            341.1        57.5         (256.2 ,488.6)
     Chao1-bc                      331.8        54.1         (251.6 ,470.2)
     ACE (Chao & Lee, 1992)        378.1        65.2         (280.7 ,543.4)
     ACE-1 (Chao & Lee, 1992)      480.3       126.2         (308.5 ,828.0)
     1st order jackknife           242.5        14.2         (218.2 ,274.4)
     2nd order jackknife           317.9        24.6         (275.9 ,373.0)
     Gamma-Poisson Model
         (Chao & Bunge, 2002)          *           *         (    * ,    *)
     Gamma-Poisson-UMLE                *           *         (    * ,    *)
     Gamma-Poisson-CMLE                *           *         (    * ,    *)
    ------------------------------------------------------------------------

     * iterative steps do not converge.


(3)  DESCRIPTION OF ESTIMATORS/MODELS:
```

(Same as in Example 1a, thus this part is omitted here.)

For this example, the estimated CV_rare = 0.686, which provides evidence of heterogeneity. The estimate (282) based on a homogeneous model should therefore severely underestimate total species richness. In this particular case, the Chao1 estimator provides a lower bound estimate of 341, while we recommend the estimate returned by the ACE estimator of 378 species with a standard error of 65.2 and a 95% confidence interval of (280.7, 543.4). For this example, the iterations in the gamma-Poisson models again failed to converge; therefore, no output is shown for them.

Example 1c (Presence/Absence Data for Multiple Samples) Seedlings Data in Data1c.txt

The seedlings data are described in Data Input Formats (3) for one community. There are 121 soil samples and 34 observed species. The running procedures are the same as those for Example 1a except that the data entry type is different.

For presence/absence data, the cut-off point again separates the data into two groups, but here we call the groups the “frequent” and “infrequent” groups (corresponding to “abundant” and “rare”

groups in the previous sections). For example, if we use a default cut-off value of 10, then those species present in at least 11 samples are placed into the "frequent" group and the other species present 1 to 10 samples are placed in the "infrequent" group. If there are less than 10 samples, then all species are placed in the infrequent group. Let Q_k (incidence count) denote the number of species that were detected in exactly k samples. Thus, only the counts $(Q_1, Q_2, ..., Q_{10})$ are used to estimate the number of undetected species. These counts along with other information for "frequent" and "infrequent" groups are displayed in the **first part** of the output shown below.

The **second part** of the output shown below lists all estimation results while the **third part** describes estimators/models used in the program. All estimators considered in the example are taken from the context of capture-recapture studies (e.g., individual animals) because there is a simple analogy between species richness estimation and population size estimation, and the species-by-sample incidence matrix is similar to the capture-recapture history matrix used for population size estimation. The capture probabilities in a capture-recapture study thus correspond to detection probabilities of species, defined here as the chance of encountering at least one individual of a given species.

Two useful models are model(h) and model(th), where h stands for "heterogeneity" and "t" stands for time/space. Model(h) assumes that the detection probabilities are heterogeneous among species while model(th) assumes that the detection probabilities vary not only among species but also among quadrats or samples (in time and/or space). Again, the CV is used to measure the degree of heterogeneity among the species detection probabilities. The sample coverage approach proposed by Lee and Chao (1994) is adopted for these two models, but the sample coverage estimator in that paper is replaced by an improved one (Shen, 2003). See the Appendix for formulas. The estimator under Model(h) is referred to as the **Incidence-based Coverage (ICE)** estimator in the ecological literature, e.g., in the program EstimateS (Colwell, 1997) and Colwell and Coddington (1994).

The estimator provided in Chao (1987) for incidence data is called the **Chao2** estimator in the program EstimateS (Colwell, 1997) and Colwell and Coddington (1994). Like the Chao1 estimator for abundance or frequency data, the Chao2 estimator was developed for multiple incidence data, and it only uses the numbers of uniques (species detected in only one sample/quadrat) and duplicates (species detected in two samples/quadrats) to estimate the number of undetected species. Using the seedlings data, all methods produce very similar point estimates of species richness, with the exception of the jackknife and beta-binomial models. The ICE estimate is 35 with a standard error of 1.0 and a 95% confidence interval of (34.2, 39.1).

We recommend to use the Chao2 and the Model(h)-1 estimate (if the estimated CV, i.e., CV_infreq in the first part of the output exceeds 0.8, then the Model(h)-1 estimator is suggested to be replaced by the Model(th)-1 estimates.) In most cases when using a default cut-off value of 10, SPADE will return estimates with the following ordering: Model(h)-1 > Model(h) > Chao2. We suggest/recommend that you use a cut-off value >10 if (a) the Model(h) estimate is lower than the Chao2 estimate, or (b) for extremely heterogeneous communities, meaning any community with very unequal detection probabilities.

Remark 1: The Chao2 estimator for $Q_2 > 0$ has the form of $\hat{S}_{Chao2} = D + [(t-1)/t]\, Q_1^2/(2Q_2)$. Like the Chao1 estimator, it returns a quite precise lower bound. The variance of this estimator is approximately $\hat{\text{var}}(\hat{S}_{Chao2}) = Q_2\left[\frac{K}{2}\left(\frac{Q_1}{Q_2}\right)^2 + K^2\left(\frac{Q_1}{Q_2}\right)^3 + \frac{1}{4}K^2\left(\frac{Q_1}{Q_2}\right)^4\right]$, where $K = (t\text{-}1)/t$. In the special case of $Q_2 = 0$, the Chao2 estimator is modified to $D + KQ_1(Q_1-1)/2$; see also Remark 2.

Remark 2: If all species have the same detection probabilities in any sample, then a "bias-corrected" estimator has the following form:

$$\hat{S}^*_{Chao2} = D + [(t-1)/t]Q_1(Q_1-1)/[2(Q_2+1)].$$

It is always obtainable, even in the case of $Q_2 = 0$. The variance formula for this bias-corrected estimator for $Q_2 > 0$ is

$$\hat{\text{var}}(\hat{S}^*_{Chao2}) = \frac{KQ_1(Q_1-1)}{2(Q_2+1)} + \frac{K^2Q_1(2Q_1-1)^2}{4(Q_2+1)^2} + \frac{K^2Q_1^2Q_2(Q_1-1)^2}{4(Q_2+1)^4}.$$

In the special case that $Q_2 = 0$, it is reduced to $\hat{S}^*_{Chao2} = D + KQ_1(Q_1-1)/2$. The variance is modified to:

$$\hat{\text{var}}(\hat{S}^*_{Chao2}) = \frac{KQ_1(Q_1-1)}{2} + \frac{K^2Q_1(2Q_1-1)^2}{4} - \frac{K^2Q_1^4}{4\hat{S}^*_{Chao2}}.$$

Similar remark as Remark 3 for abundance data apply to presence/absence data as well.

OUTPUT:

```
(1)  BASIC DATA INFORMATION:

          (Number of observed species) D =    34
          (Number of samples/quadrats) t =   121
          (Cut-off point)              k =    10

     "Infrequent" Species Group (Incidence counts up to the cut-off point)
        ---------------------------------------------------------------------
        Q[1]=3; Q[2]=2; Q[3]=3; Q[4]=3; Q[5]=1; Q[6]=5; Q[7]=1;
        Q[8]=1; Q[9]=3; Q[10]=1;
        ---------------------------------------------------------------------

       (Number of observed species for infrequent species)  D_infreq  = 23
       (Estimated sample coverage for infrequent species)   C_infreq  = 0.974
       (Estimated CV for infrequent species)                CV_infreq = 0.371

     "Frequent" Species Group: (Incidence counts over the cut-off point)
       (Number of observed species for frequent species)      D_freq = 11

(2) ESTIMATION OF SPECIES RICHNESS:
```

Estimator/Model	Estimate	Est_s.e.	95% Confidence interval	Est. CV (rare)
Homogeneous Model	34.6	0.9	(34.1, 38.9)	
Chao2 (Chao, 1987)	36.2	3.4	(34.3, 52.9)	
Chao2-bc	35.0	2.3	(34.1, 48.0)	
Model(h) or ICE	35.0	1.0	(34.2, 39.1)	0.371
Model(h)-1 or ICE-1	35.1	1.0	(34.2, 39.2)	0.398
Model(th)	35.1	1.4	(34.2, 41.4)	0.387
Model(th)-1	35.1	1.5	(34.2, 42.0)	0.415
1st order jackknife	37.0	2.4	(34.7, 46.1)	
2nd order jackknife	38.0	4.2	(34.7, 55.7)	
Beta-binomial-CMLE	44.2	11.5	(35.7, 94.4)	
Beta-binomial-UMLE	41.4	7.9	(35.3, 74.9)	

(3) DESCRIPTION OF ESTIMATORS/MODELS:

Homogeneous Model: This model assumes that all species have the same detection probabilities. See Eq.(3.2) of Lee and Chao (1994).

Chao2 (Chao, 1987): This approach uses the frequencies of uniques and duplicates to estimate the number of missing species; see Chao (1987).

Chao2-bc: a bias-corrected form for the Chao2; see Chao (2005).

Model(h) (ICE: Incidence-based Coverage Estimator): Model(h) assumes that the detection probabilities are heterogeneous among species. The estimator given here is an improved version of Eq.(3.18) in Lee and Chao (1994) by using an improved estimated sample coverage given in Shen (2003) and the SPADE User Guide; see Eq.(3.23) in Lee and Chao (1994) for the estimated squared CV.

Model(h)-1 (or ICE-1): A modified ICE for highly-heterogeneous cases.

Model(th): Model(th) assumes that the detecting probability varies not only among species but also among quadrats/samples. The estimator given here is an improved version of Eq.(3.18) in Lee and Chao (1994) by using an improved estimated sample coverage given in Shen (2003) and the SPADE User Guide; see Eq.(3.22) in Lee and Chao (1994) for the estimated squared CV.

Model(th)-1: A modified estimator for highly-heterogeneous cases under Model(th).

1st order jackknife: It uses the frequency of uniques to estimate the number of missing species; see Burnham and Overton (1978).

2nd order jackknife: It uses the frequencies of uniques and duplicates to estimate the number of missing species; see Burnham and Overton (1978).

Beta-Binomial-CMLE: Conditional MLE under the beta-binomial model in which the number of samples of any species is detected follows a binomial distribution with probability being a random variable from a beta distribution.

Beta-Binomial-UMLE: Unconditional MLE under the beta-binomial model.

95% Confidence interval: A log-transformation is used so that the lower bound of the resulting interval is at least the number of observed species. See Chao (1987).

Example 1d (Frequencies Data for Multiple Samples) Seedlings Data in Data1d.txt

The detailed description for this data set is given in Data Input Formats (4) for one community. All the steps are the same as those in Example 1c. Note that there are two approaches to this type of data. The first approach is to convert the abundance matrix to an incidence matrix, and then to use a similar analysis as that in Example 1c. This approach ignores the abundance counts and is thus reduced to Example1c. The second approach is to sum all abundance counts (or frequencies) for any fixed species over the multiple samples and apply the procedures mentioned in Example 1a. The output based on the abundance data for the seedlings example in Data1d.txt is shown below (see Example 1a for the description of the output). Note that the ACE estimator in the example shown below and the ICE estimator in Example 1c yield very similar results. Actually, all estimates shown in the output below are quite similar.

Suggested guideline If your multiple abundance data were collected over the same area using a similar method, then it is suggested to sum all abundance counts for each species; otherwise it is suggested to convert your abundance data to incidence data especially for quadrat sampling.

OUTPUT:

```
(1)  BASIC DATA INFORMATION:

        (Number of observed species)     D = 34
        (Number of samples/quadrats)     t = 121
        (Number of observed individuals) n = 952
        (Cut-off point)                  k = 10

    Rare Species Group: (Frequency counts up to the cut-off point)
        ---------------------------------------------------------------------------
         f[1]=2; f[2]=2; f[3]=1; f[4]=4; f[6]=2; f[7]=1; f[8]=1;
         f[9]=1; f[10]=1;
        ---------------------------------------------------------------------------
        (Number of observed individuals for rare species)     n_rare  =    71
        (Number of observed species for rare species)         D_rare  =    15
        (Estimated sample coverage for rare species)          C_rare  =  0.97
        (Estimated CV for rare species)                       CV_rare =  0.42

    Abundant Species Group: (Frequencies beyond the cut-off point)
        (Number of observed individuals for abundant species) n_rare =   881
        (Number of observed species for abundant species)     D_rare =    19

(2)  Estimation Results of the Number of Species:

    Estimator/Model            Estimate    Est_s.e.  95% Confidence interval
    ------------------------------------------------------------------------
    Homogeneous Model             34.4        0.7        (  34.0,  38.3)
    Homogeneous (MLE)             34.0        0.0        (  34.0,  34.0)
    Chao1 (Chao, 1984)            35.0        1.9        (  34.1,  45.1)
    Chao1-bc                      34.3        0.9        (  34.0,  40.0)
    ACE (Chao & Lee, 1992)        34.8        1.2        (  34.1,  40.7)
    ACE-1 (Chao & Lee, 1992)      34.8        1.3        (  34.1,  41.3)
    1st order jackknife           36.0        2.0        (  34.4,  44.2)
    2nd order jackknife           36.0        3.5        (  34.2,  54.1)
    Gamma-Poisson Model
        (Chao & Bunge, 2002)      34.8        3.2        (  34.0,  55.7)
    Gamma-Poisson-UMLE               *          *        (     *,     *)
```

```
Gamma-Poisson-CMLE              *          *        (     *,     *)
-------------------------------------------------------------------------
* iterative steps do not converge.
```

(3) DESCRIPTION OF ESTIMATORS/MODELS:

(Same as in Example 1a, thus this part is omitted here.)

Example 1e (Incidence Counts Data for Multiple Samples) Seedlings Data in Data1e.txt

The seedlings data and output are the same as those in Example 1a; the only difference is that the data entry type is different in this case. Here, the required data type is described in Data Input Format (5) (Incidence Counts Data for Multiple Samples/Quadrats).

Part II: Shared Species (Estimating Shared Spceies Richness in Two Communities)

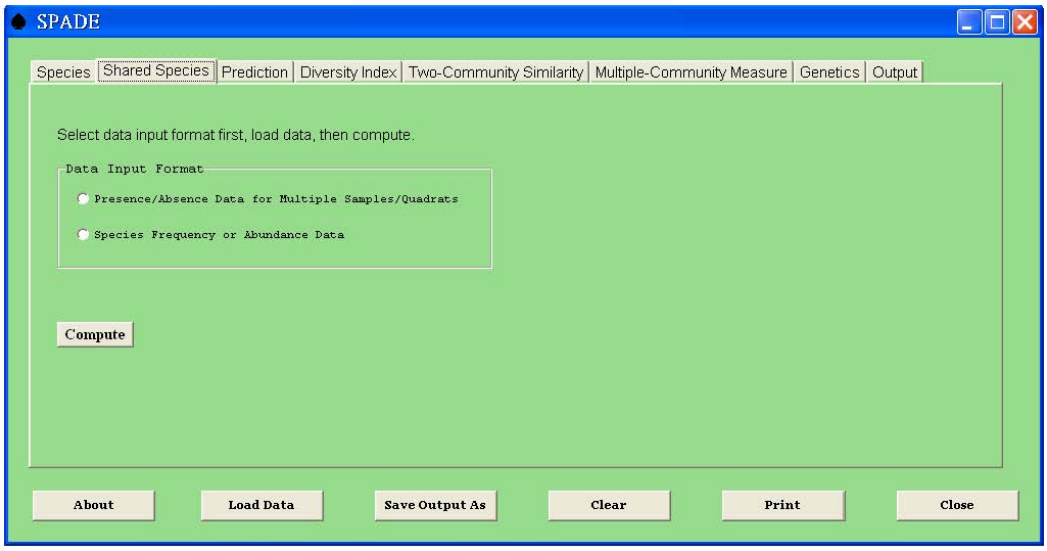
This part of the program features estimates for the number of shared species in two communities based on the following two sampling schemes:

(1) In each community, a random sample of individuals is taken and species frequencies or abundances are recorded;

(2) Each community is sampled several times or the whole area is divided into several quadrats and species presence/absence data for multiple samples/quadrats are recorded.

The reader is referred to Chao et al. (2000) and Chao et al. (2006b) for the models and theoretical backgrounds. When you select "Shared Species" in the top-level menu, the following window is shown:

Figure 2: The Window of SPADE for Estimating Shared Species Richness



There are two types of data input formats corresponding to the above two sampling schemes.

(1) Species Frequency or Abundance Data: This type is described in Data Input Formats (1) for two communities. An example using **data2a.txt** is given below.
(2) Presence/absence Data for Multiple Samples/Quadrats: This type is described in Data Input Formats (2) for two communities. An example using **data2b.txt** is given below.

Example2a: Birds Abundance Data in Two Estuaries: (Data2a.txt)

This example with data stored in **Data2a.txt** was analyzed in details by Chao et al. (2000). A total of 85867 and 59646 observations have been made from two estuaries (which are hereafter referred to as Community I and II respectively). In these two communities, there were, respectively, 155 and 140 species observed, with 111 of these recorded for both areas (shared species). The purpose was to estimate the true number of shared species in the two communities because it is expected that some shared species were not observed or only observed in one community. The step-by-step running procedures are the following:

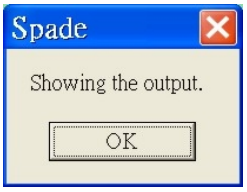
Step 1. Select proper data type from Data Structure.

Step 2. Load your data by pressing the key Load Data, then input the path to the data file.

Step 3. Then enter the number of bootstrap replications in order to construct a confidence interval. The default value is 200.

Press OK, the screen is minimized as shown below.

Step 4. The computation is finished when the following window is shown:



The output prompts out under Output Menu for your browse. You can then press the button Save Output As to save it in a designated file or press Print to get a hard copy.

The output shown far below includes five parts. The first part shows basic data information for the two samples. The cut-off point separates the observed shared species into two groups: rare and abundant. Among the observed 111 shared species, there were 21 rare shared species which were observed less than or equal to 10 times in both communities; these are classified to "rare" shared species group and the other 90 shared species are classified to "abundant" group. Only the former group is used for estimating the number of undetected shared species. For those 21 species in the "rare" group, the sample sizes and relevant information are also shown. The second part shows four estimates of shared species proposed in Chao et al. (2000) and Chao et al. (2006b); see below. The third part provides description of models used in the second part. The fourth part provides four species richness estimators (Chao1, Chao1-bc, ACE and ACE-1) for

each single community and the fifth part gives description of those four estimators. See Example 1a for description of the four estimators.

The homogeneous model assumes that the shared species in each community have the same discovery probabilities. This model yields an estimate of 114 (s.e. 7.4) for the number of shared species. In practice, the homogeneous model is rarely valid. The ACE and Chao1 estimators for estimating species richness in one community have been respectively extended to estimate shared species richness for two communities. The corresponding shared species richness estimators for ACE, Chao1 and Chao1-bc (bias-corrected of Chao1) are named as ACE-shared, Chao1-shared and Chao1-shared-bc in the output.

For the ACE-shared estimator, we use coefficient of covariation (CCV) to characterize the heterogeneity of species discovery probabilities in each community and the correlation of the two sets of the discovery probabilities; see Chao et al. (2000, pp. 232-233). Some relevant formulas are provided in the Appendix. As shown in the output, the CCV estimates are $CCV_1 = 0.7328$, $CCV_2 = 1.0072$ and $CCV_{12} = 0.4574$; see Equations (3.9a) and (3.9b) of Chao et al. (2000). These large values of CCVs show strong evidence for heterogeneity and correlation exist. Therefore, we need to incorporate the estimated CCVs in the resulting estimator. The estimated number of shared species in the ACE approach by Chao et al. (2000) is 134. A bootstrap s.e. estimate based on 200 bootstrap replications is approximately 21.3, which implies a 95% confidence interval of (110, 201). Therefore, we can conclude that there are still 23 shared species not discovered in the survey; a 95% confidence interval for the true number of shared species is (111, 201). We remark that the bootstrap resampling procedures vary with trial, thus two different runs may result in different s.e. estimates. This quite long confidence interval signifies that shared data are not sufficient for modeling heterogeneous communities. Similar findings hold for both the Chao1-shared and Chao1-shared-bc estimators.

Chao, Shen and Hwang (2006) extended the Chao1 estimator for a single community to the two-community case and provided an estimator of shared species richness. Like the Chao-1 estimator, the Chao1-shared estimator is very simple and Chao1-shared is always obtainable. Unlike the Chao1, the Chao1-shared cannot be theoretically proved to be a lower bound.

OUTPUT:

```
(1)  BASIC DATA INFORMATION:

             (Number of observed individuals in community 1) n1  =85867
             (Number of observed individuals in community 2) n2  =59646
             (Number of observed species in community 1)     D1  =155
             (Number of observed species in community 2)     D2  =140
             (Number of observed shared species)             D12 =111
             (Bootstrap replications for s.e. estimate)      10


    "Rare" Shared Species Group: (Both Frequencies can only up to 10)
        Some Statistics:
        ------------------------------------------------------------
         f[11] = 4; f[1+] = 8; f[+1] = 9; f[2+] = 1; f[+2] = 4
        ------------------------------------------------------------
            (Number of observed individuals in community 1) n1_rare  = 3358
            (Number of observed individuals in community 2) n2_rare  = 558
```

```
(Number of observed shared species)          D12_rare = 21
(Estimated sample coverage)                  C12_rare = 0.8596
(Estimated CCVs)                             CCV_1    = 0.7328
                                             CCV_2    = 1.0072
                                             CCV_12   = 0.4574
```

(2) ESTIMATION RESULTS OF THE NUMBER OF SHARED SPECIES:

Model	Estimate	Est_s.e.	95% CI (percentile)
Homogeneous	114.43	7.4	(111.0 ,128.3)
Heterogeneous (ACE-shared)	133.92	21.3	(111.0 ,201.4)
Chao1-shared	162.79	38.0	(125.3 ,298.7)
Chao1-shared-bc	145.38	22.3	(121.8 ,220.7)

- for the case C12_rare=0 (ACE-shared is replaced by Chao1-shared)
* for the case f[2+]=0 or f[+2]=0 (Chao1-shared is replaced by Chao1-shared-bc)

(3) DESCRIPTION OF MODELS FOR ESTIMATING SHARED SPECIES RICHNESS:

Homogeneous: This model assumes that the shared species in each community have the same discovery probabilities; see the Eq.(3.11a) of Chao et al. (2000).

Heterogeneous (ACE-shared): This model allows for heterogeneous discovery probabilities among shared species; see Eq.(3.11b) of Chao et al. (2000). It is an extension of ACE to two communities. It is replaced by Chao1-shared when the estimated sample coverage (C12_rare in the output) is zero.

Chao1-shared: An extension of the Chao1 estimator to two communities. It provides an estimate of shared species richness. See Eq. (11) of Chao, Shen and Hwang (2006). It is replaced by Chao1-shared-bc (see below) for the case f[2+]=0 or f[+2]=0.

Chao1-shared-bc: A bias-corrected form of the Chao1-shared. See Eq. (12) of Chao, Shen and Hwang (2006).

(4) SINGLE COMMUNITY ANALYSIS:

Model	Estimate	Est_s.e.	95% CI
Community 1:			
Chao1	207.08	30.6	(172.9 ,306.6)
Chao1-bc	197.86	24.0	(170.4 ,274.1)
ACE	181.34	11.2	(166.9 ,213.4)
ACE-1	192.83	18.4	(170.3 ,248.3)
Community 2:			
Chao1	160.17	11.3	(147.2 ,196.2)
Chao1-bc	157.77	10.0	(146.3 ,189.8)
ACE	162.15	9.6	(149.8 ,190.1)
ACE-1	170.88	15.1	(152.4 ,216.7)

(5) DESCRIPTIONS OF MODELS FOR SINGLE COMMUNITY ANALYSIS:

(This part is the same as Output Part (3) in Example 1a.)

Example2b: The Hong Kong Big Bird Race (BBR) : (Data2b.txt)

This data set is described in the Data Input Formats (2) in two communities and was analyzed in details by Chao, Shen and Hwang (2006). In 1999, a total of 217 species was observed by 19 teams competed in Hong Kong BBR. In 2000, a total of 220 species was observed. Merging the two-year data by species names, we found that there were 115 observed shared species. In 1999, the winning team recorded 152 species. This means that the winning team missed 65 species that were observed by at least one of the other teams. In 2000, the winning team recorded 154 species; thus, 66 species that were observed by the other teams were missed by the winning team. The purpose of analyzing these data basically was the following: First, were there any species missed by all teams in each race? Second, did the data provide sufficient evidence that there are more species in 2000 than in 1999? Finally, were there any unobserved shared species in the two sets of data? The step-by-step running procedures are the same as in example 2a except for the data type is different. The output is shown below. There are five parts in the output, and these five parts include similar types of results or descriptions except that in the second part, we only have the Chao2-shared and Chao2-shared-bc. (The ICE method has not been extended to the shared species richness estimation yet.)

We first present the species richness estimation for one community. In the fourth part of output, four species richness estimators (Chao2, Chao2-bc, ICE and ICE-1) based on multiple incidence data are given; see Example 1c for description. The output shows that in 1999 the Chao2 estimate is 234 (s.e. 10.2) and whereas in 2000 the corresponding Chao2 estimate is 233 (s.e. 7.5). The ICE gives an estimate of 229 (s.e. 3.3) for 1999, and of 232 (s.e. 3.4) for 2000. As a result, there appears insufficient evidence to support that there are more bird species in 2000 than in 1999 in Hong Kong.

From the merged (by species identity) data of the two years, we obtain frequency counts $Q_{1+} = 6$, $Q_{2+} = 4$, $Q_{+1} = 10$, $Q_{+2} = 7$ and $Q_{11} = 1$, which are the main statistics used in Chao2-shared. Based on the Chao2-shared estimate, the shared species richness is 127 (s.e. 7.6) with a 95% confidence interval of (119, 153). Thus, we can conclude that there were about 12 shared species not detected with a 95% confidence interval of (3, 38). See Chao, Shen and Hwang (2006) doe detailed analysis.

OUTPUT:

```
(1)  BASIC DATA INFORMATION:

                    (Number of samples from community 1)   t1=    19
                    (Number of samples from community 2)   t2=    20
              (Number of observed species in community 1)   D1=   217
              (Number of observed species in community 2)   D2=   220
    (Number of observed shared species in two communities) D12=   115


    Some Statistics:
    ------------------------------------------------
     Q[11]=1; Q[1+]=6; Q[+1]=10; Q[2+]=4; Q[+2]=7
    ------------------------------------------------

(2) ESTIMATION RESULTS OF THE NUMBER OF SHARED SPECIES:

    Model             Estimate   Est_s.e.        95% CI
   ------------------------------------------------------------
    Chao2-shared        126.53       7.6     (118.5, 152.6)
```

```
    Chao2-shared-bc     123.52       6.2     (117.4, 145.3)
   ---------------------------------------------------------

(3) DESCRIPTION OF MODELS FOR ESTIMATING SHARED SPECIES RICHNESS:

    Chao2-shared: An extension of the Chao2 estimator to two communities. It provides an estimate of shared
        species richness. See Eq. (13) of Chao, Shen and Hwang (2006). It is replaced by Chao2-shared-bc
        (see below) for the case Q[2+]=0 or Q[+2]=0.

    Chao2-shared-bc:  A bias-corrected form of Chao2-shared. See Eq. (14) of Chao, Shen and Hwang (2006).

(4) SINGLE COMMUNITY ANALYSIS:

     Model           Estimate    Est_s.e.           95% CI
    ------------------------------------------------------------
     Community 1
       Chao2          234.22       10.2        (222.9 ,267.3)
       Chao2-bc       232.00        9.2        (222.0 ,262.4)
       ICE            228.46        3.3        (223.6 ,237.0)
       ICE-1          230.52        3.7        (225.0 ,239.7)
     Community 2
       Chao2          233.09        7.5        (224.6 ,257.3)
       Chao2-bc       231.74        7.0        (224.0 ,254.8)
       ICE            231.93        3.4        (226.9 ,240.5)
       ICE-1          233.90        3.7        (228.4 ,243.1)
    ------------------------------------------------------------

(5) DESCRIPTIONS OF MODELS FOR SINGLE COMMUNITY ANALYSIS:
```

(This part is the same as Output Part (3) in Example 1c.)

Part III: Prediction (Predicting the Number of New Species in a Further Survey)

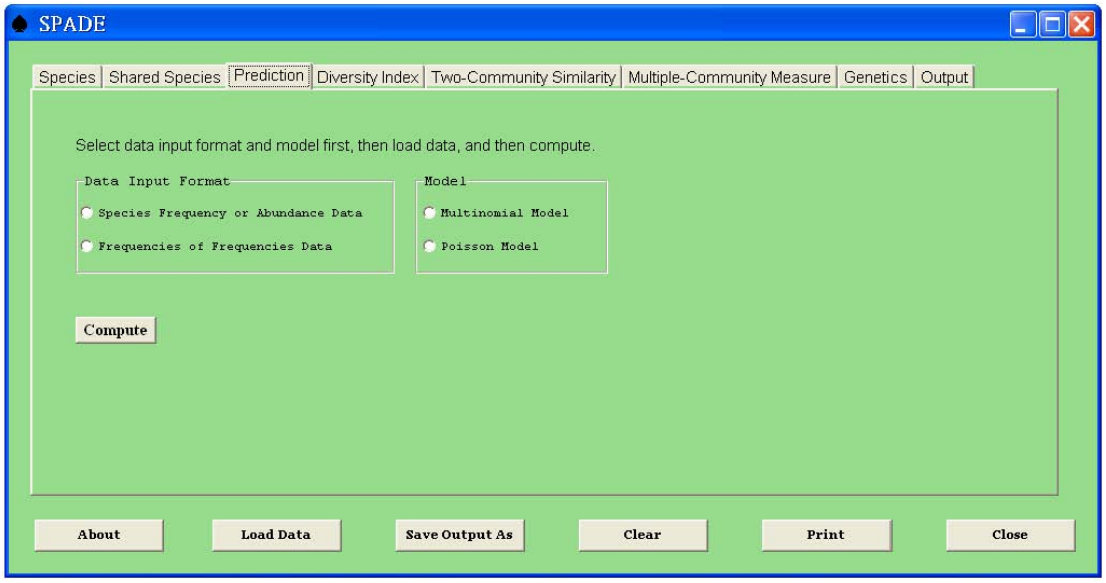
This part provides prediction results for the number of new species that would be discovered in a second survey, based on data from an initial survey. The window for Prediction Menu is shown below (Figure 3). There are two models for selections:

(1) The Multinomial Model: (Shen, Chao and Lin, 2003)
This model is applicable when a sample of n individuals is taken from a community. The goal is to predict the number of new species in a second survey of m individuals.

(2) The Poisson Model (Chao and Shen, 2004)
This model is applicable when the community is sampled for a fixed time of interval or fixed amount of efforts. Without losing of generality, it is assumed to be one unit of time interval or unit of effort for the original survey. The goal is to predict the number of new species in an additional time interval of t units.

There are two data input formats (1) species frequency or abundance data and (2) frequencies of frequencies data. See Data Input Formats for details on data entry.

Figure 3. The Window of SPADE for Prediction.



Example 3a (Frequencies of frequencies) Plant Data in Data3a.txt

We use the data on rare vascular plant species tabulated in Miller and Wiegert (1989) as an example. A total of 188 species were recorded out of n = 1008 individuals compiled over a 150-year period. The frequency counts of species frequency are reproduced in the following table.

Table 3.1. Frequency Counts of the Extant Rare Plant Species (Miller and Wiegert, 1989)

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
f_i	61	35	18	12	15	4	8	4	5	5	1	2	1	2	3

i	16	19	20	22	29	32	40	43	48	67	Species seen	#Individuals
f_i	2	1	2	1	1	1	1	1	1	1	188	1008

The data are stored in **Data3a.txt** in the format of frequencies of frequencies. The data are read as

67 25 1 61 2 35 3 18 4 12 ...48 1 67 1

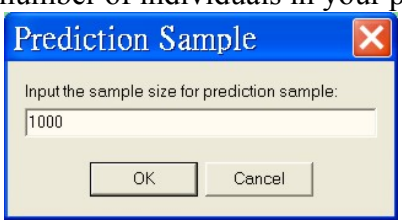
Here the first entry 67 denotes the maximum frequency and the second entry 25 denotes the number of frequencies with non-zero counts. Then followed by 25 pairs (1, 61), (2, 35) ...(67, 1). Suppose that we are interested in estimating the number of new species expected in the next m = 1000 additional individuals. The running procedures are the follows:

Step 1. Select "Multinomial Model" from Model menu.

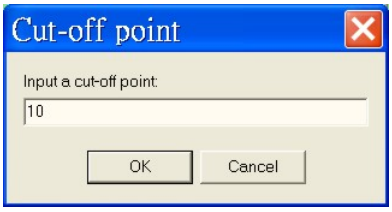
Step 2. Select "Frequencies of Frequencies Data" from Data Input Format.

Step 3. Load your data by pressing the key Load Data, then select the complete path of the data file.

Step 4. Press the key Compute to calculate estimates. The following window prompts out asking for the entry of the number of individuals in your prediction sample.



Step 5. In the computing procedure, an estimate of species richness is needed, thus a cut-off point is needed as those procedure in Part I for estimating species richness. (default cut-off = 10)



Press OK, the window screen is minimized to

The computation is complete when the following window prompts out:

Step 6. Press OK, and the output prompts out under Output Menu for your browse. You can then press the button Save Output As to save it in a designated file or press Print to get a hard copy.

OUTPUT:

```
(1)  BASIC DATA INFORMATION:

        (Number of observed individuals)  n = 1008
        (Number of observed species)      D = 188
        (Prediction size)                 m = 1000
        (Cut-off point)                   k = 10
        (Estimated sample coverage)       C = 0.939

     Rare Species Group: (Frequency counts up to the cut-off point)
           ---------------------------------------------------------------------
            f[1]=61; f[2]=35; f[3]=18; f[4]=12; f[5]=15; f[6]=4; f[7]=8;
            f[8]=4; f[9]=5; f[10]=5;
           ---------------------------------------------------------------------
           (Number of observed individuals for rare species)  n_rare  = 515
           (Number of observed species for rare species)      D_rare  = 167
           (Estimated sample coverage for rare species)       C_rare  = 0.882
           (Estimated CV for rare species)                    CV_rare = 0.715

     Abundant Species Group: (frequencies beyond the cut-off point)
           (Number of observed individuals for abundant species)  n_abun = 493
           (Number of observed species for abundant species)      D_abun = 21

(2)  PREDICTION OF THE NUMBER OF NEW SPECIES IN A FURTHER SURVEY:

     Estimator/Model             Estimate     Est_s.e.    95% Confidence interval
    ------------------------------------------------------------------------------
     Efron & Thisted (1976)          43.8         12.9     (      18.5 ,      69.06)
```

```
Boneh et al. (1998)              24.1        2.1     (      19.9 ,      28.3)
Solow & Polasky (1999)           36.1        7.7     (      21.1 ,      51.2)
Shen et al. (2003)               37.5        6.9     (      24.1 ,      51.0)
-------------------------------------------------------------------------------
For a review of these four estimators, see Shen et al. (2003).


(3)  ESTIMATION OF SPECIES RICHNESS:

Estimator/Model          Estimate    Est_s.e.       95% Confidence interval
-------------------------------------------------------------------------------
Chao1 (Chao, 1984)          241.2       17.9           ( 216.0,  288.9  )
Chao1-bc                    238.8       17.1           ( 214.7,  284.6  )
ACE (Chao & Lee, 1992)      245.8       15.2           ( 222.8,  284.0  )
ACE-1 (Chao & Lee, 1992)    265.4       22.7           ( 232.0,  324.0  )
-------------------------------------------------------------------------------
See Example 1a for description of estimators in Output Part (3).
```

The output includes three parts. The first part presents data basic information. The second part shows the prediction or estimation results for four estimators/models proposed respectively by Efron and Thisted (1976), Boneh et al. (1998), Solow and Polasky (1999) and Shen et al. (2003). See Shen et al. (2003) for a review of these four methods and see the Appendix for relevant formulas. The third part shows four estimators of the species richness based on the initial sample. The first and third parts are similar to those in Examples 1a and 1b.

From the second part output, the Shen et al. method predicts that in a further survey of 1000 individuals there would be 38 new species with 95% confidence interval of (24, 51). The estimated species richness in the third part is needed in the method by Solow and Polasky (in which the Chao1 is adopted) and by Shen et al. (in which the ACE is adopted). This estimate in the third part can be used to compute the estimated number of unseen species in the original sample, i.e., the asymptotic value as the prediction size m tends to infinity. For example, as the prediction size becomes large, Solow and Polasky (1999) estimates would approach to 53 (= 241 – 188) whereas Shen et al. (2003) estimates tend to 58 (= 246 -188).

When the prediction size is larger than the initial size, the estimate by Efron and Thisted may become extremely large and useless. For example, if in Step 2, we enter 2000 as prediction size, then the results for the second part are shown below, where the Efron and Thisted estimate becomes extremely large.

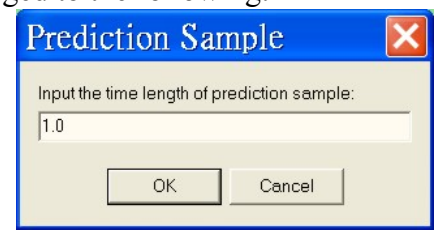
```
(2)  PREDICTION OF THE NUMBER OF NEW SPECIES IN A FURTHER SURVEY:

Estimator/Model          Estimate     Est_s.e.   95% Confidence interval
-------------------------------------------------------------------------------
Efron & Thisted (1976)   8.7E+019    8.6E+019    ( -8.3E+019 ,  2.6E+020)
Boneh et al. (1998)          30.7         2.4     (      26.0 ,      35.4)
Solow & Polasky (1999)       47.7        12.1     (      24.0 ,      71.4)
Shen et al. (2003)           50.7         9.9     (      31.2 ,      70.2)
-------------------------------------------------------------------------------
For a review of these four estimators, see Shen et al. (2003).
```

Example 3b (Frequencies of frequencies) Butterfly Data in Data3b.txt

For a Poisson model, we use the Malayan butterfly data analyzed in Fisher et al. (1943) for an illustrative example. In this data set, a total of 620 species were observed out of 9031 butterflies. All frequency counts are given in Williams (1964, p.19). Assume that the sampling time of the initial survey is 1, the goal here is to estimate the number of new species expected in an additional time interval of the same length. The frequency counts are stored in **Data3b.txt** and read as

194 75 1 118 2 74 3 44 4 24 5 29 … 194 1

The data in all steps from Step 2 to Step 6 are the same as those in Example 3a except that the window in Step 4 is changed to the following:



The results under Poisson model and prediction time interval $t = 1$ are shown below.

OUTPUT:

```
(1)  BASIC DATA INFORMATION:

        (Number of observed individuals)          n = 9031
        (Number of observed species)              D = 620
        (The time length of prediction sample)    t = 1.000
        (Cut-off point)                           k = 10
        (Estimated sample coverage)               C = 0.987

     Rare Species Group: (Frequency counts up to the cut-off point)
           ---------------------------------------------------------------------
            f[1]=118; f[2]=74; f[3]=44; f[4]=24; f[5]=29; f[6]=22; f[7]=20;
            f[8]=19; f[9]=20; f[10]=15;
           ---------------------------------------------------------------------
           (Number of observed individuals for rare species)  n_rare  = 1393
           (Number of observed species for rare species)      D_rare  = 385
           (Estimated sample coverage for rare species)       C_rare  = 0.915
           (Estimated CV for rare species)                    CV_rare = 0.662

     Abundant Species Group: (frequencies beyond the cut-off point)
           (Number of observed individuals for abundant species)  n_abun = 7638
           (Number of observed species for abundant species)      D_abun = 235

(2)  PREDICTION OF THE NUMBER OF NEW SPECIES IN A FURTHER SURVEY:

     Estimator/Model          Estimate    Est_s.e.     95% Confidence interval
    ---------------------------------------------------------------------------
     Efron & Thisted (1976)       78.0        24.7     (      29.5 ,      126.5)
     Boneh et al. (1998)          49.0         2.3     (      44.4 ,       53.5)
     Chao & Shen (2004)           66.5         7.8     (      51.2 ,       81.9)
    ---------------------------------------------------------------------------
     For a review of these three estimators, see Chao and Shen (2004).

(3)  ESTIMATION OF SPECIES RICHNESS:

          Estimator            Estimate     Est_s.e.          95% Confidence interval
    -------------------------------------------------------------------------------
```

Chao1 (Chao, 1984)	714.1	22.7	(679.1, 769.9)
Chao1-bc	712.0	22.2	(677.7, 766.7)
ACE (Chao & Lee, 1992)	712.1	17.3	(683.9, 752.8)
ACE-1 (Chao & Lee, 1992)	737.2	27.4	(694.6, 804.2)

Under a Poisson model, only three estimators are considered. The method by Solow and Polasky (1999) method is specifically derived only under a multinomial model. Therefore, it is not included for the Poisson model. The method by Chao and Shen (2004) indicates that if an additional sample with approximately the same size as the original sample is conducted, then the predicted number of new species would be 66 with a 95% confidence interval of (51, 82).

In the third part, there are four estimators; Chao and Shen (2004) adopt the ACE estimate. As in Example 3a, this ACE can be used to compute the number of unseen species in the original sample, that is, the asymptotic value as the prediction interval length *t* tends to infinity. For example, as the prediction interval becomes large, the Chao and Shen (2004) estimates would approach to 92 (= 712-620).

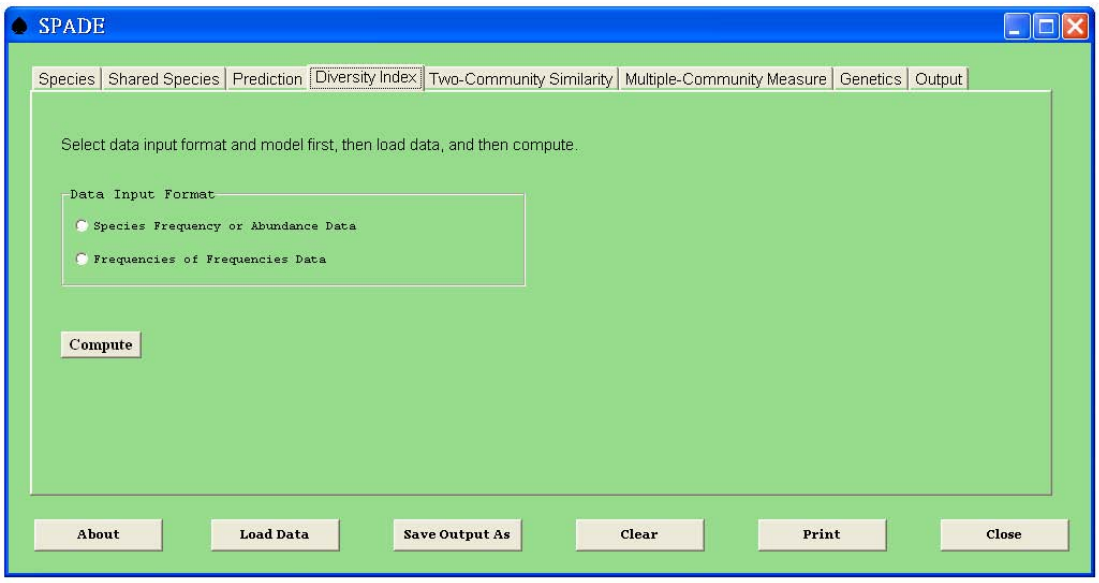
Remark: When you apply both multinomial and Poisson models to the same set of frequency counts, the two models will result in slightly different results. For examples, there is small discrepancy between the two estimated CV and predicted results are somewhat different. These slight discrepancies are due to model differences.

Part IV: Diversity Index (Estimating Various Diversity Indices)

This part features various diversity indices including the Shannon's index and its effective number of species (true diversity of order 1), the Simpson's index and its effective number of species (true diversity order 2), species richness (true diversity of order 0), and the Fisher's alpha index. See Chao and Shen (2003) for a review of various estimators for Shannon's index and Jost (2006, 2007) for the concept and background for true (or neutral) diversity. Relevant formulas are given in the Appendix.

There are two data input formats (1) species frequency or abundance data and (2) frequencies of frequencies data. See Data Input Formats for details on data entry. The window for this part is shown below.

Figure 4. The Interface of SPADE for Estimating Diversity Index.



Example 4a: Tree Frequency Data in Data4a.txt

The data stored in **Data4a.txt** include the sample tree frequencies from an old-growth rain forest named LEP. The original data were collected by Dr. Robin Chazdon and colleagues and discussed in Chao et al. (2008).

Running Procedures:

Step 1. Select proper data type from Data Input Formats in the menu.
(For example 4a, select Species Frequency or Abundance Data)

Step 2. Load your data by pressing the key Load Data and input the complete path of the data file.

Step 3. Press the key Compute to calculate estimates. Then the window screen is minimized to

The computation is complete when the following window pops up:

The output is shown below (see the following example 4b for interpretation of the output):

```
(1)  BASIC DATA INFORMATION:

     (Number of observed individuals)      n  = 557
     (Number of observed species)          D  = 69
     (Estimated sample coverage)           C  = 0.957
     (Estimated CV)                        CV = 2.237

(2)  ESTIMATION OF SPECIES RICHNESS (TRUE DIVERSITY OF ORDER 0):
```

Estimator	Estimate	Est_s.e.	95% Confidence interval
Chao1 (Chao, 1984)	105.0	20.3	(81.8, 170.0)
Chao1-bc	99.7	16.9	(80.1, 153.3)
ACE (Chao & Lee, 1992)	92.1	10.2	(79.1 , 121.8)
ACE-1 (Chao & Lee, 1992)	100.4	15.7	(81.4, 148.1)

95% Confidence interval: A log-transformation is used so that the lower bound of the resulting interval is at least the number of observed species. See Chao (1987).

Chao1 (Chao, 1984): This approach uses the numbers of singletons and doubletons to estimate the number of missing species because missing species information is mostly concentrated on those low frequency counts; see Chao (1984), Shen, Chao and Lin (2003) and Chao, Shen and Hwang (2006).

Chao1-bc: a bias-corrected form for the Chao1; see Chao (2005).

ACE (Abundance-based Coverage Estimator): A non-parametric estimator proposed by Chao and Lee (1992) and Chao, Ma and Yang (1993). The observed species are separated as rare and abundant groups; only the rare group is used to estimate the number of missing species. The estimated CV is used to characterize the degree of heterogeneity among species discovery probabilities. See Eq.(2.14) in Chao and Lee (1992) or Eq.(2.2) of Chao et al. (2000).

ACE-1: A modified ACE for highly heterogeneous communities; See Eq.(2.15) of Chao and Lee (1992).

(3a) SHANNON INDEX:

Estimator	Estimate	Est_s.e.	95% Confidence interval
MLE	3.193	0.060	(3.076, 3.311)
MLE_bc	3.275	0.210	(2.864, 3.686)
Jackknife	3.280	0.066	(3.150, 3.410)
Chao & Shen	3.308	0.201	(2.913, 3.702)

For a review of the four estimators, see Chao and Shen (2003).

MLE: maximum likelihood estimator.
MLE_bc: bias-corrected maximum likelihood estimator.
Jackknife: see Zahl (1977).
Chao & Shen: based on Horvitz-Thompson estimator and sample coverage method; see Chao and Shen (2003).

(3b) EXPONENTIAL OF SHANNON INDEX (TRUE DIVERSITY OF ORDER 1):

Estimator	Estimate	Est_s.e.	95% Confidence interval
MLE	24.372	1.461	(21.508,27.235)
MLE_bc	26.449	5.548	(15.576,37.322)
Jackknife	26.573	1.759	(23.118,30.028)
Chao & Shen	27.320	5.496	(16.548,38.092)

(4a) SIMPSON INDEX:

Estimator	Estimate	Est_s.e.	95% Confidence interval
MVUE	0.08328	0.01797	(0.04807, 0.11850)

```
    MLE           0.08493         0.01791      (0.04983, 0.12002  )
   ------------------------------------------------------------------

    MVUE: minimum variance unbiased estimator; see Eq. (2.27) of Magurran (1988).
    MLE: maximum likelihood estimator; see Eq. (2.26) of Magurran (1988).

(4b)  INVERSE OF SIMPSON INDEX (TRUE DIVERSITY OF ORDER 2):

    Estimator    Estimate        Est_s.e.     95% Confidence interval
   ------------------------------------------------------------------
    MVUE         12.00729         0.21574      (11.58445, 12.43013  )
    MLE          11.77460         0.21083      (11.36138, 12.18783  )
   ------------------------------------------------------------------

(5)  FISHER ALPHA INDEX:

                 Estimate        Est_s.e.     95% Confidence interval
   ------------------------------------------------------------------
    alpha          20.739          2.497        (15.845, 25.632  )
   ------------------------------------------------------------------

    See Eq. (2.9) of Magurran (1988) for a definition of Fisher's alpha index.
```

Example 4b: (Frequencies of frequencies) Insects data in Data4b1.txt and Data4b2.txt

To illustrate our method, we selected two data sets from Janzen (1973a, b) when he collected tropical foliage insects. The following table gives the frequencies of frequencies for beetles collected respectively in day-time and night-time from the site referred to as "Osa primary-hill, dry season, 1967" in Janzen's paper (1973a). The day-time and night-time data are stored in **Data4b1.txt** and **Data4b2.txt**, respectively.

Table 4.1. Frequency Counts for Insects Data

Day-time (in file Data4b1.txt)

m	1	2	3	4	5	6	11	species seen	individuals
f_m	59	9	3	2	2	2	1	78	127

Night-time (in file Data4b2.txt)

m	1	2	3	5	7	10	14	16	18	species seen	individuals
f_m	56	9	7	2	1	1	1	1	1	79	170

OUTPUT (for Day-time data):

```
(1)  BASIC DATA INFORMATION:

    (Number of observed individuals)      n  = 127
    (Number of observed species)          D  = 78
    (Estimated sample coverage)           C  = 0.535
    (Estimated CV)                        CV = 1.207

(2)  ESTIMATION OF SPECIES RICHNESS (TRUE DIVERSITY OF ORDER 0):
```

Estimator	Estimate	Est_s.e.	95% Confidence interval
Chao1 (Chao, 1984)	271.4	83.0	(164.4, 510.8)
Chao1-bc	249.1	69.5	(157.6, 445.9)
ACE (Chao & Lee, 1992)	263.3	64.4	(173.6 , 437.1)
ACE-1 (Chao & Lee, 1992)	415.4	146.7	(227.3, 840.6)

95% Confidence interval: A log-transformation is used so that the lower bound of the resulting interval is at least the number of observed species. See Chao (1987).

Chao1 (Chao, 1984): This approach uses the numbers of singletons and doubletons to estimate the number of missing species because missing species information is mostly concentrated on those low frequency counts; see Chao (1984), Shen, Chao and Lin (2003) and Chao, Shen and Hwang (2006).

Chao1-bc: a bias-corrected form for the Chao1; see Chao (2005).

ACE (Abundance-based Coverage Estimator): A non-parametric estimator proposed by Chao and Lee (1992) and Chao, Ma and Yang (1993). The observed species are separated as rare and abundant groups; only the rare group is used to estimate the number of missing species. The estimated CV is used to characterize the degree of heterogeneity among species discovery probabilities. See Eq.(2.14) in Chao and Lee (1992) or Eq.(2.2) of Chao et al. (2000).

ACE-1: A modified ACE for highly heterogeneous communities; See Eq.(2.15) of Chao and Lee (1992).

(3a) SHANNON INDEX:

Estimator	Estimate	Est_s.e.	95% Confidence interval
MLE	4.077	0.074	(3.932, 4.222)
MLE_bc	5.110	0.375	(4.374, 5.845)
Jackknife	4.624	0.110	(4.406, 4.842)
Chao & Shen	4.699	0.211	(4.287, 5.112)

For a review of the four estimators, see Chao and Shen (2003).

MLE: maximum likelihood estimator.
MLE_bc: bias-corrected maximum likelihood estimator.
Jackknife: see Zahl (1977).
Chao & Shen: based on Horvitz-Thompson estimator and sample coverage method; see Chao and Shen (2003).

(3b) EXPONENTIAL OF SHANNON INDEX (TRUE DIVERSITY OF ORDER 1):

Estimator	Estimate	Est_s.e.	95% Confidence interval
MLE	58.971	4.351	(50.442,67.500)
MLE_bc	165.625	62.143	(43.825,287.426)
Jackknife	101.906	11.217	(79.709,124.104)
Chao & Shen	109.869	23.129	(64.536,155.202)

(4a) SIMPSON INDEX:

Estimator	Estimate	Est_s.e.	95% Confidence interval

```
MVUE          0.01687         0.00459      (0.00788, 0.02586  )
MLE           0.02461         0.00436      (0.01607, 0.03315  )
-----------------------------------------------------------------
```

MVUE: minimum variance unbiased estimator; see Eq. (2.27) of Magurran (1988).
MLE: maximum likelihood estimator; see Eq. (2.26) of Magurran (1988).

(4b) INVERSE OF SIMPSON INDEX (TRUE DIVERSITY OF ORDER 2):

```
Estimator    Estimate        Est_s.e.     95% Confidence interval
-----------------------------------------------------------------
MVUE         59.26667         0.27180      (58.73394, 59.79940  )
MLE          40.62720         0.17701      (40.28027, 40.97414  )
-----------------------------------------------------------------
```

(5) FISHER ALPHA INDEX:

```
             Estimate        Est_s.e.     95% Confidence interval
-----------------------------------------------------------------
alpha          86.009          9.739        (66.921, 105.096  )
-----------------------------------------------------------------
```

See Eq. (2.9) of Magurran (1988) for a definition of Fisher's alpha index.

****

OUTPUT (for Night-time data):

(1) BASIC DATA INFORMATION:

```
(Number of observed individuals)      n  = 170
(Number of observed species)          D  = 79
(Estimated sample coverage)           C  = 0.671
(Estimated CV)                        CV = 1.714
```

(2) ESTIMATION OF SPECIES RICHNESS (TRUE DIVERSITY OF ORDER 0):

```
          Estimator       Estimate    Est_s.e.       95% Confidence interval
-------------------------------------------------------------------------------
Chao1 (Chao, 1984)           253.2        75.6             ( 156.2,  472.2  )
Chao1-bc                     233.0        63.4             ( 149.9,  413.3  )
ACE (Chao & Lee, 1992)       268.5        69.7             ( 173.3 , 460.1  )
ACE-1 (Chao & Lee, 1992)     471.9       206.2             ( 228.4, 1112.3  )
-------------------------------------------------------------------------------
```

95% Confidence interval: A log-transformation is used so that the lower bound
of the resulting interval is at least the number of observed species.
See Chao (1987).

Chao1 (Chao, 1984): This approach uses the numbers of singletons and doubletons to
estimate the number of missing species because missing species information is
mostly concentrated on those low frequency counts; see Chao (1984), Shen, Chao and Lin (2003)
and Chao, Shen and Hwang (2006).

Chao1-bc: a bias-corrected form for the Chao1; see Chao (2005).

ACE (Abundance-based Coverage Estimator): A non-parametric estimator proposed by Chao and Lee (1992)
and Chao, Ma and Yang (1993). The observed species are separated as rare and abundant groups;

only the rare group is used to estimate the number of missing species.
The estimated CV is used to characterize the degree of heterogeneity among species discovery probabilities. See Eq.(2.14) in Chao and Lee (1992) or Eq.(2.2) of Chao et al. (2000).

ACE-1: A modified ACE for highly heterogeneous communities; See Eq.(2.15) of Chao and Lee (1992).

(3a) SHANNON INDEX:

Estimator	Estimate	Est_s.e.	95% Confidence interval
MLE	3.832	0.088	(3.661, 4.004)
MLE_bc	4.619	0.377	(3.881, 5.357)
Jackknife	4.236	0.118	(4.003, 4.469)
Chao & Shen	4.297	0.205	(3.895, 4.700)

For a review of the four estimators, see Chao and Shen (2003).

MLE: maximum likelihood estimator.
MLE_bc: bias-corrected maximum likelihood estimator.
Jackknife: see Zahl (1977).
Chao & Shen: based on Horvitz-Thompson estimator and sample coverage method; see Chao and Shen (2003).

(3b) EXPONENTIAL OF SHANNON INDEX (TRUE DIVERSITY OF ORDER 1):

Estimator	Estimate	Est_s.e.	95% Confidence interval
MLE	46.161	4.040	(38.242,54.079)
MLE_bc	101.387	38.185	(26.544,176.229)
Jackknife	69.125	8.151	(53.034,85.215)
Chao & Shen	73.511	15.098	(43.919,103.104)

(4a) SIMPSON INDEX:

Estimator	Estimate	Est_s.e.	95% Confidence interval
MVUE	0.03341	0.00623	(0.02120, 0.04563)
MLE	0.03910	0.00605	(0.02724, 0.05096)

MVUE: minimum variance unbiased estimator; see Eq. (2.27) of Magurran (1988).
MLE: maximum likelihood estimator; see Eq. (2.26) of Magurran (1988).

(4b) INVERSE OF SIMPSON INDEX (TRUE DIVERSITY OF ORDER 2):

Estimator	Estimate	Est_s.e.	95% Confidence interval
MVUE	29.92708	0.18656	(29.56142, 30.29275)
MLE	25.57522	0.15478	(25.27186, 25.87859)

(5) FISHER ALPHA INDEX:

	Estimate	Est_s.e.	95% Confidence interval
alpha	57.367	6.454	(44.717, 70.018)

See Eq. (2.9) of Magurran (1988) for a definition of Fisher's alpha index.

For both examples 4a and 4b, the output is divided into five parts: Output (1) includes basic data information: the sample size, observed species richness, estimated sample coverage and CV. Output (2) provides four species estimators (true diversity of order 0); these four estimators are extracted from the output of Part I which features other estimators; see Examples 1a and 1b for details. Output (3a) shows four estimates of Shannon's entropy index; its effective number of species (true diversity of order 1) is given in Output (3b). Output (4a) shows two estimates of Simpson's index; its effective number of species (true diversity of order 2) is given in Output (4b). Output (5) provides the estimate of Fisher's alpha index.

We use Example 4b for illustration. For these two data sets, most species had only one, two or three individuals represented in the sample, with only a few abundant species. Therefore, the data information is concentrated in the lower-order capture frequencies. The estimated CV values for the day-time and night-time data are, respectively, 0.938 and 1.099, based on the first ten frequency counts $\{f_1, f_2, ..., f_{10}\}$. You may run Part I to look at these values using a cut-off = 10 as the default input value. Note that in the output the value for day-time data is CV = 1.207 and for night-time data is CV = 1.714 in the above output of this part analysis refers to all observed species including rare and abundant ones. These relatively high CV values indicate that the community is highly heterogeneous in species abundances; therefore, any species richness estimator based on a homogeneous model that does not incorporate the heterogeneity would have severe negative bias. The guidelines in Example 1a suggest the use of the ACE-1 estimator for both data sets. However, their standard errors are too large to make them useful point estimates. Thus, we recommend to use the Chao1 estimator or the ACE estimator; these two estimates of species richness imply that a relatively large fraction of species have remained undetected in the samples.

For the estimation of Shannon's index, a comparison of four estimates and their standard errors is made in Table 4.2. This table shows that the traditional maximum likelihood estimator (MLE) has the lowest estimate, but the bias-corrected MLE yields the highest estimate. The jackknife and Chao and Shen (2003) estimates are in-between, but are still higher than the MLE estimate. The bias-corrected MLE has the lowest estimated precision. All estimates imply that the diversity of the day-time data is higher than that of the night-time data. Although the estimated standard error for the Chao and Shen (2003) estimator is larger than that of the jackknife estimator, and therefore the associated confidence interval is greater. Chao and Shen (2003) concluded that their method produces a confidence interval with coverage probability being closer to the nominal level.

Table 4.2. Comparison of Various Estimates of Shannon's Index of Diversity for Janzen's Insects Data (estimated s.e. are in parentheses)

Estimate	Day-time	Night-time
MLE	4.08 (0.07)	3.83 (0.09)
Bias-corrected MLE	5.11 (0.38)	4.62 (0.38)
Jackknife	4.62 (0.11)	4.24 (0.12)

Chao and Shen (2003)	4.70 (0.21)	4.30 (0.21)

Table 4.3 gives the estimated effective numbers of species (true diversity) of orders 0, 1 and 2 for both day-time and night-time data. For the total species richness based on the ACE and Chao1 estimators, the relatively high standard errors signify difficulties in obtaining precise estimates for these two data sets. However, the true diversity of orders 1 and 2 clearly show that the day-time community is more diverse than the night-time community.

Table 4.3. Comparison of True Diversity of Orders 0, 1 and 2
(estimated s.e. are in parentheses)

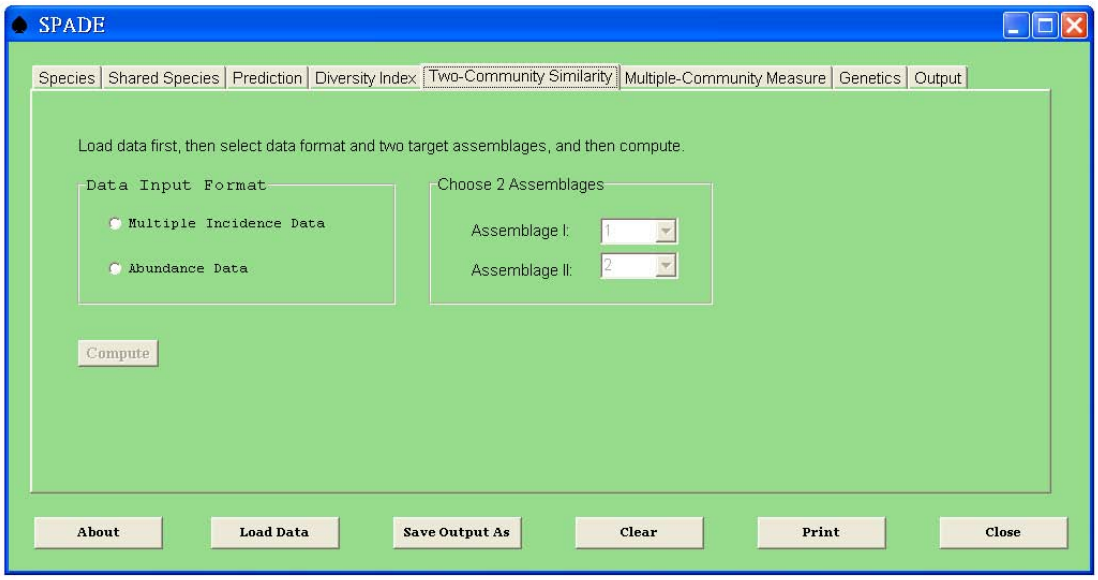
Estimate	Day-time	Night-time
"True" diversity of order 0 (ACE)	263 (64.4)	269 (69.7)
"True" diversity of order 0 (Chao1)	271 (83.0)	253 (75.6)
"True" diversity of order 1 (Chao and Shen, 2003)	110 (23.1)	74 (15.1)
"True" diversity of order 2 (MVUE)	59 (0.27)	30 (0.19)

Part V: Two-Community Similarity Index

This part features various similarity indices for comparing data from two assemblages based on either multiple-sample incidence data or abundance data. The incidence-based indices include the classic Jaccard, Sørensen and Lennon et al. (2001) indices, and the abundance-based indices include the Bray-Curtis, Morisita-Horn and two newly developed abundance-based Jaccard and Sørensen indices. See Chao et al. (2005) for a review of various indices and estimation details and Chao et al. (2006a) for relative merits of all indices.

There are two data input formats (1) multiple incidence data and (2) abundance data. See Data Input Formats for details on data entry. The window for this part is shown below.

Figure 5. The Interface of SPAE for Estimating Two-Community Similarity Indices



You may enter data for more than two assemblages. For examples, the three assemblages in Example 5a are species frequencies for trees, saplings and seedlings. This part only provides similarity indices for two assemblages. The extension to indices for more than two assemblages is featured in Part VI.

Running Procedures:
(Please note in this part, you must load data first.)

Step 1. Load your data by pressing the key Load Data, then input the complete path of the data file.
Step 2. Select proper data type from Data Input Formats in the menu.
Step 3. Select two target assemblages from your data for comparison.
Step 4. Press the key Compute to calculate estimates.

Example 5a: (Abundance data) Rain Forest Data in Data5a.txt

This data set is described in Data Output Formats for the case of more than two assemblages. The data stored in **Data5a.txt** include the frequencies from three assemblages: seedlings (column 1), saplings (column 2) and trees (column 3) in an old-growth rain forest named LEP. Suppose we like to assess the similarity between seedlings and trees, select "Abundance Data" first in the Window menu and also select Assemblage 1 and Assemblage 3 as the two target assemblages. (The three assemblages will be indexed by 1, 2 and 3 according to your ordering of data entry.) Then load data and compute all various similarity indices discussed in Chao et al. (2005). The output for comparing seedlings vs. trees is shown below.

Remark 1: The s.e. estimates in the following output are different from those in the paper by Chao et al. (2005, 2006a) because we have modified our original bootstrap variance estimation procedures. Our original variance estimates over-estimate in many applications. The current modified procedure yield more reasonable s.e. estimates.

Remark 2: The diversity analysis within each community is featured in Part IV (Diversity Index). The true diversity of orders 0, 1 and 2 as well as their s.e. estimate and 95% confidence interval are provided in that part. Please run Part IV for single-community diversity analysis. Please refer to example 4a in Part IV for interpretation of output. See Jost (2006, 2007, 2008) for theoretical backgrounds.

OUTPUT:

```
(1) The loaded set includes abundance (or frequency) data from 3 assemblages
    (indexed by 1, 2, 3 according to data entry order) and a total of 120 distinct species.

(2) The two chosen assemblages for comparison: 1 vs. 3

(3) Basic Data Information:

        (Number of observed individuals in Assemblage  1)  n1=   557
        (Number of observed individuals in Assemblage  3)  n2=   111
            (Number of observed species in Assemblage  1)  D1=    69
            (Number of observed species in Assemblage  3)  D2=    43
   (Number of observed shared species in two assemblages) D12=    26
                    (Bootstrap replications for s.e. estimate)   200

    Some Statistics:
         ------------------------------------------------------------
          f[11] = 4; f[1+] = 8; f[+1] = 10; f[2+] = 2; f[+2] = 8
         ------------------------------------------------------------

(4) Estimation Results of Some Similarity Indices:

                                  Estimate   Bootstrap se.     U_hat* (se.)     V_hat** (se.)
    ------------------------------------------------------------------------------------------
    Incidence-based:
    ==============
      Jaccard incidence               0.3023       0.0281
      Sorensen incidence              0.4643       0.0360
      Lennon et al (2001)             0.6047       0.0532

    Abundance-based:
    ==============
      Bray-Curtis                     0.2425       0.0158
      Morisita-Horn                   0.7442       0.0412
      Morisita Original               0.7868       0.0413
      Jaccard Abundance  (unadjusted) 0.4040       0.0268        0.4578             0.7748
      Jaccard Abundance  (  adjusted) 0.4845       0.0498        0.4845 (0.0574)    1.0000 (0.0940)
      Sorensen Abundance (unadjusted) 0.5755       0.0286        0.4578             0.7748
      Sorensen Abundance (  adjusted) 0.6527       0.0475        0.4845 (0.0574)    1.0000 (0.0940)
    ------------------------------------------------------------------------------------------
      *  U denotes the total relative abundances of the shared species in the first assemblage;
           U_hat is an estimate of U.
      ** V denotes the total relative abundances of the shared species in the second assemblage;
           V_hat is an estimate of V.

(5) References:

    Chao, A., Chazdon, R. L., Colwell, R. K. and Shen, T.-J. (2005). A new statistical approach for assessing
```

similarity of species composition with incidence and abundance data. Ecology Letters, 8, 148-159.

Chao, A., Chazdon, R. L., Colwell, R. K. and Shen, T.-J. (2006). Abundance-based similarity indices and their estimation when there are unseen species in samples. Biometrics, 62, 361-371.

(6) ESTIMATION RESULTS OF THE NUMBER OF SPECIES FOR EACH ASSEMBLAGE:

Model	Estimate	Est_s.e.	95% CI
Assemblage 1:			
Chao1	105.00	20.3	(81.8 ,170.0)
Chao1-bc	99.67	16.9	(80.1 ,153.3)
ACE	92.13	10.2	(79.1 ,121.8)
ACE-1	100.38	15.7	(81.4 ,148.1)
Assemblage 2:			
Chao1	69.45	14.8	(52.5 ,116.4)
Chao1-bc	66.00	12.8	(51.3 ,106.5)
ACE	74.03	14.7	(55.8 ,118.2)
ACE-1	85.11	24.0	(57.9 ,162.0)

(7) DESCRIPTIONS OF MODELS FOR SINGLE ASSEMBLAGE ANALYSIS:

Chao1 (Chao, 1984): This approach uses the numbers of singletons and doubletons to estimate the number of missing species because missing species information is mostly concentrated on those low frequency counts; see Chao (1984), Shen, Chao and Lin (2003) and Chao, Shen and Hwang (2006).

Chao1-bc: a bias-corrected form for the Chao1; see Chao (2005).

ACE (Abundance-based Coverage Estimator): A non-parametric estimator proposed by Chao and Lee (1992) and Chao, Ma and Yang (1993). The observed species are separated as rare and abundant groups; only the rare group is used to estimate the number of missing species. The estimated CV is used to characterize the degree of heterogeneity among species discovery probabilities. See Eq. (2.14) in Chao and Lee (1992) or Eq.(2.2) of Chao et al. (2000).

ACE-1: A modified ACE for highly heterogeneous communities; See Eq.(2.15) of Chao and Lee (1992).

Example 5b: (Multiple incidence data) Ant data in Data5b.txt

This data set is described in Data Output Formats for the case of more than two assemblages. The data include incidence frequencies of tropical rainforest ants using three techniques: (a) Berlese extraction of soil samples (217 samples) (b) fogging samples from canopy fogging (459 samples), and (c) Malaise trap samples for flying and crawling insects (62 samples). For comparing the two methods (Fogging vs. Malaise), we select “multiple incidence data” and two assemblages (2 and 3) in the Window Menu, then comes the following output: (See Chao et al. (2005) for interpretation.)

Remark: The s.e. estimates in the following output are different from those in the paper by Chao et al. (2005, 2006a) because we have modified our original bootstrap variance estimation procedures. Our original variance estimates over-estimate in many applications. The current modified procedure yield more reasonable s.e. estimates.

OUTPUT:

(1) The loaded set includes multiple-sample incidence data from 3 assemblages

(indexed by 1, 2, 3 according to data entry order) and a total of 276 distinct species

(2) The two chosen assemblages for comparison: 2 vs. 3

(3) Basic Data Information:

(Number of samples from assemblage 2) w=	459
(Number of samples from assemblage 3) z=	62
(Number of observed species in assemblage 2) D1=	165
(Number of observed species in assemblage 3) D2=	103
(Number of observed shared species in two assemblages) D12=	78
(Bootstrap replications for s.e. estimate)	200

Some Statistics:

Q[11] = 3; Q[1+] = 6; Q[+1] = 25; Q[2+] = 2; Q[+2] = 14

(4) Estimation Results of Some Similarity Indices:

	Estimate	Bootstrap. se.	U_hat* (se.)	V_hat** (se.)
Incidence-based:				
Jaccard incidence	0.4105	0.0167		
Sorensen incidence	0.5821	0.0180		
Lennon et al (2001)	0.7573	0.0219		
Multiple incidence-based:				
Bray-Curtis	0.1872	0.0090		
Morisita-Horn	0.6333	0.0322		
Morisita Original	0.6679	0.0339		
Incidence-based Jaccard (unadjusted)	0.6504	0.0209	0.7410	0.8418
Incidence-based Jaccard (adjusted)	0.7476	0.0378	0.8285 (0.0477)	0.8845 (0.0276)
Incidence-based Sorensen (unadjusted)	0.7881	0.0165	0.7410	0.8418
Incidence-based Sorensen (adjusted)	0.8556	0.0273	0.8285 (0.0477)	0.8845 (0.0276)

* U denotes the total relative incidences of the shared species in the first assemblage; U_hat is an estimate of U.

** V denotes the total relative incidences of the shared species in the second assemblage; V_hat is an estimate of V.

(5) References:

Chao, A., Chazdon, R. L., Colwell, R. K. and Shen, T.-J. (2005). A new statistical approach for assessing similarity of species composition with incidence and abundance data. Ecology Letters, 8, 148-159.

Chao, A., Chazdon, R. L., Colwell, R. K. and Shen, T.-J. (2006). Abundance-based similarity indices and their estimation when there are unseen species in samples. Biometrics, 62, 361-371.

(6) ESTIMATION RESULTS OF THE NUMBER OF SPECIES FOR EACH ASSEMBLAGE:

Model	Estimate	Est_s.e.	95% CI
Assemblage 1			
Chao2	180.72	9.0	(170.6 ,209.5)
Chao2-bc	178.97	8.3	(169.8 ,206.1)
ICE	177.62	3.4	(172.5 ,186.3)

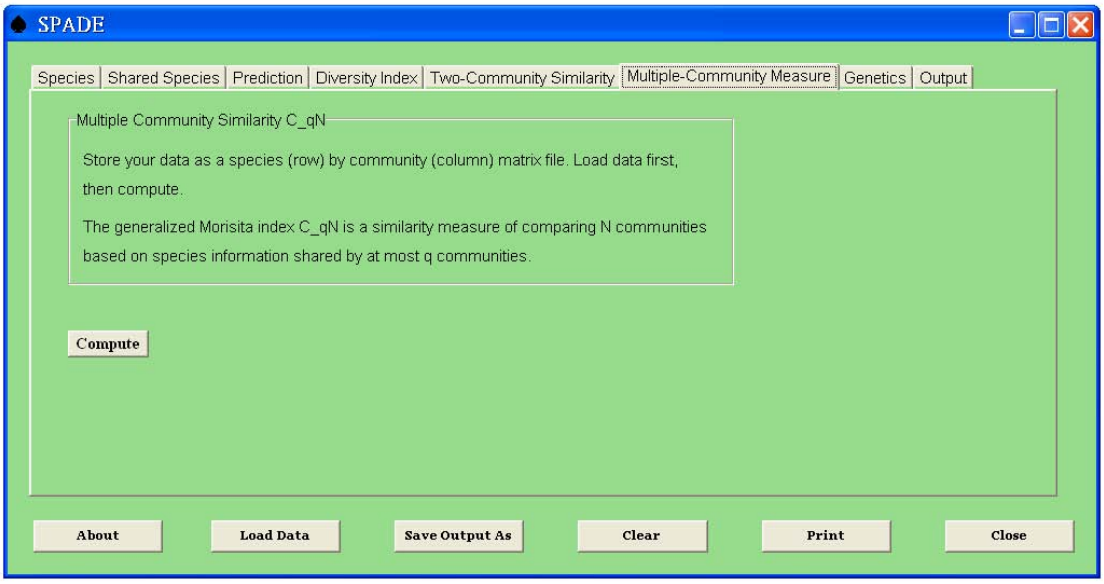
```
   ICE-1          179.80       3.8        (174.0 ,189.2)
 Assemblage 2
   Chao2          144.57      17.7        (121.6 ,195.7)
   Chao2-bc       141.37      16.6        (120.1 ,189.3)
   ICE            150.58       6.6        (139.3 ,165.4)
   ICE-1          172.13      11.2        (153.5 ,197.7)
-------------------------------------------------------------

(7) DESCRIPTIONS OF MODELS FOR SINGLE ASSEMBLAGE ANALYSIS:
    Chao2 (Chao, 1987): This approach uses the frequencies of uniques and duplicates to
        estimate the number of missing species; see Chao (1987).
    Chao2-bc: a bias-corrected form for the Chao2; see Chao (2005).
    ICE (Incidence-based Coverage Estimator): Model(h) assumes that the detection
        probabilities are heterogeneous among species. The estimator given here is an improved
        version of Eq.(3.18) in Lee and Chao (1994) by using an improved estimated sample coverage
        given in Shen (2003) and the SPADE User Guide; see Eq.(3.23) of Lee and Chao (1994)
        for the estimated squared CV.
    ICE-1:  A modified ICE for highly-heterogeneous cases.
```

Part VI: Multiple-Community Diversity Measure

This part computes the generalized Morisita similarity/dissimilarity indices for comparing frequency or abundance data from more than two communities. A traditional approach for simultaneously comparing more than two communities in ecology is to use multiple pairwise comparisons. However, Chao et al. (2008) gave an example to show that two sets of communities have the same pairwise similarity but they are not globally similar. Thus, pairwise similarity indices do not completely characterize multiple-community similarity because the information shared by at least three communities is ignored.

For comparing N communities, a profile of $N-1$ indices is proposed to characterize similarity of species composition across communities. The profile includes the indices C_{2N}, C_{3N}, …, $C_{N\,N}$, where C_{jN} is an overall similarity index based on the information shared by at most j communities. In this part, nearly unbiased estimators of the multiple-community Morisita indices and their variances are obtained. The interface window for this part is shown below.

Figure 6: The Interface Window of SPADE for Multiple-Community Measure



Remark: For single-community diversity analysis, run Part IV (Diversity Analysis) and refer to example 4a in Part IV for interpretation of output.

Running Procedures:
Store your data as a species (row) by community (column) matrix file.

Step 1. Load your data by pressing the key Load Data, then input the complete path of the data file.

Step 2. Press the key Compute to calculate estimates.

Example 6a: Rain Forest Data (Abundance Data)

The data stored in **Data6a.txt** (same as Data5a.txt) include the frequencies from three assemblages: seedlings (column 1), saplings (column 2) and trees (column 3) in the LEP old-growth rain forest. In Example 5a, only the first and the third columns are compared. In this part we illustrate the similarity indices for comparing three communities. After loading Data6a.txt and computing, the output for comparing these three assemblages is shown below.

```
(1) BASIC DATA INFORMATION:

      The loaded set includes abundance (or frequency) data from 3 communities
      and a total of 120 distinct species.


      (Number of observed individuals in each community)        n1=557
                                                               n2=729
                                                               n3=111

      (Number of observed species in one community)             D1=69
                                                               D2=102
                                                               D3=43

      (Number of observed shared species in two communities)    D12=59
                                                               D13=26
                                                               D23=32

      (Number of observed shared species in three communities)  D123=23

      (Bootstrap replications for s.e. estimate)               200

(2) ESTIMATION OF MORISITA SIMILARITY IN 3 COMMUNITIES

      Estimator   Estimate    Est_s.e.    95% Confidence Interval_1     95% Confidence Interval_2
    ------------------------------------------------------------------------------------------------------
      C23         0.613       0.026       (0.563, 0.664)                (0.556, 0.668)
      C33         0.498       0.043       (0.414, 0.581)                (0.399, 0.577)
    ------------------------------------------------------------------------------------------------------
    C23: A similarity measure of comparing 3 communities based on shared information between any two communities.
    C33: A similarity measure of comparing 3 communities using all shared information.

    Confidence Interval_1: Based on estimate (+/-) 1.96 bootstrap s.e.
    Confidence Interval_2: Based on an improved bootstrap percentile method. (recommend for use in the case when
                           similarity is close to 0 or 1 )
```

```
    Pairwise Comparison:
    ----------------------------------------------------------------------------------------------------------------
      C22(1,2)    0.471       0.037       (0.399, 0.543)              (0.399, 0.541)
      C22(1,3)    0.787       0.041       (0.706, 0.868)              (0.705, 0.864)
      C22(2,3)    0.483       0.072       (0.342, 0.624)              (0.352, 0.642)
    ----------------------------------------------------------------------------------------------------------------
    Average Pairwise =  0.580
    ----------------------------------------------------------------------------------------------------------------

    Similarity Matrix
    ----------------------------------------------------------------------------------------------------------------
    C22(i,j)  1     2     3
      1     1.000 0.471 0.787
      2           1.000 0.483
      3                 1.000
    ----------------------------------------------------------------------------------------------------------------

(3) ESTIMATION OF MORISITA DISSIMILARITY IN 3 COMMUNITIES

      Estimator   Estimate    Est_s.e.    95% Confidence Interval_1     95% Confidence Interval_2
    ----------------------------------------------------------------------------------------------------------------
      1-C23       0.387       0.026       (0.336, 0.437)              (0.332, 0.444)
      1-C33       0.502       0.043       (0.419, 0.586)              (0.423, 0.601)
    ----------------------------------------------------------------------------------------------------------------
    1-C23: This is the genetic diversity measure D defined in Jost (2008) for
           comparing 3 communities.
    1-C33: A genetic diversity measure for comparing 3 subpopulations based on
           all shared information.

    Pairwise Comparison:
    ----------------------------------------------------------------------------------------------------------------
      1-C22(1,2)  0.529       0.037       (0.457, 0.601)              (0.459, 0.601)
      1-C22(1,3)  0.213       0.041       (0.132, 0.294)              (0.136, 0.295)
      1-C22(2,3)  0.517       0.072       (0.376, 0.658)              (0.358, 0.648)
    ----------------------------------------------------------------------------------------------------------------
    Average Pairwise =  0.420
    ----------------------------------------------------------------------------------------------------------------
    1-C22: This is the genetic diversity measure D defined in Jost (2008) for
           comparing 2 subpopulations.

    Dissimilarity Matrix
    ----------------------------------------------------------------------------------------------------------------
    1-C22(i,j)    1     2     3
      1         0.000 0.529 0.213
      2               0.000 0.517
      3                     0.000
    ----------------------------------------------------------------------------------------------------------------


References:
Chao, A., Jost, L., Chiang, S. C., Jiang, Y.-H. and Chazdon, R. (2008). A Two-
    stage probabilistic approach to multiple-community similarity indices.
    Biometrics, 64, 1178-1186.
Jost, L. (2008). GST and its relatives do not measure differentiation. Molecular
    Ecology, 17, 4015-4026.
```

There are three parts in the output. First part shows some basic data information including the number of individuals and the number of species in each community; and the number of species shared by exactly two communities and the number of species shared by exactly three communities. For comparing more than three communities, only shared information between any two communities is given.

The second part shows the multiple-community Morisita similarty indices C_{23} and C_{33} and their s.e. estimates along with 95% confidence intervals. The pairwise similarity indices and their average are also given. The corresponding dis-similarity index (one minus similarity index) output follows. The s.e. estimate is obtained by a bootstrap method based on 200 replications; see Chao et al. (2008) for details.

There are two methods for constructing confidence intervals:

(1) Confidence Interval_1: Based on estimate ± 1.96 bootstrap s.e.
When an estimating target parameter is not near the boundary of parameter space (i.e., 0 or 1), the asymptotic normality of an estimate is typically valid, so a symmetric 95% CI based on a normal critical point 1.96 can be applied.

(2) Confidence Interval_2: Based on an improved bootstrap percentile method.
When a parameter is near boundary 0 or 1, then the asymptotic normality is no longer valid, (because the central limit theory used for proving asymptotic normality requires a condition of the existence of a derivative around the true parameter), and the bootstrap distribution is generally skewed. In this case, we have an asymmetric confidence interval. Thus, this method is especially recommended for use in the case when the parameter is close to 0 or 1.) Here it is hard to define "how close is close" and may be data-dependent, but this method is always valid for all parameters. If the parameter is not near any boundary, the two methods give very close results. Therefore, we would recommend the use of this method.

For this example, the pairwise Morisita index between the first and the third columns, $C_{22}(1,3)$ = 0.787, implying the similarity between trees and seedlings is high. The similarity between trees and saplings, $C_{22}(1,2)$ = 0.471 and the similarity between saplings and trees, $C_{22}(2,3)$ = 0.483 are moderate and comparable. We remark that the Morisita index is mainly sensitive to dominant species, thus the above finding emphasizes the relationships among the more abundant species in the assemblages.

The three-community Morisita index consists of a profile of two indices (C_{23}, C_{33}). Intuitively, we can regard C_{23} as an overall measure of comparing 3 communities based on shared information between any two communities; C_{33} is the global measure taking into account the proportion of individuals that belong to species shared by two and by all three communities. The estimated C_{23} for our example is 0.613, but the global measure C_{33} has a lower value of 0.498. This means that if species shared by all three communities are considered, then the global similarity is lower. A two-stage probabilistic interpretation of C_{qN} provided in Chao et al. (2008) is as follows. In Stage I, we select q out of the N communities with replacement and with equal probabilities. In Stage II, we select an individual from each of the q communities that are selected in Stage I. Then C_{qN} can be expressed as the ratio of two conditional probabilities, namely, $C_{qN} = P(Z \mid A^c)/P(Z \mid A)$, where Z denotes the event that the q selected individuals in Stage II belong to the same species, A denotes that the q selected communities in Stage I are the same, and A^c denotes the complement of A (i.e., at least one community that is different from the others).

For the case of N ($N > 3$) communities, the current version only features C_{2N} and all pairwise comparisons. The measure $1-C_{2N}$ is the differentiation measure proposed by Jost (2008). The measures C_{3N}, C_{4N}, …, $C_{N,N}$ are still under programming and will be featured in SPADE.

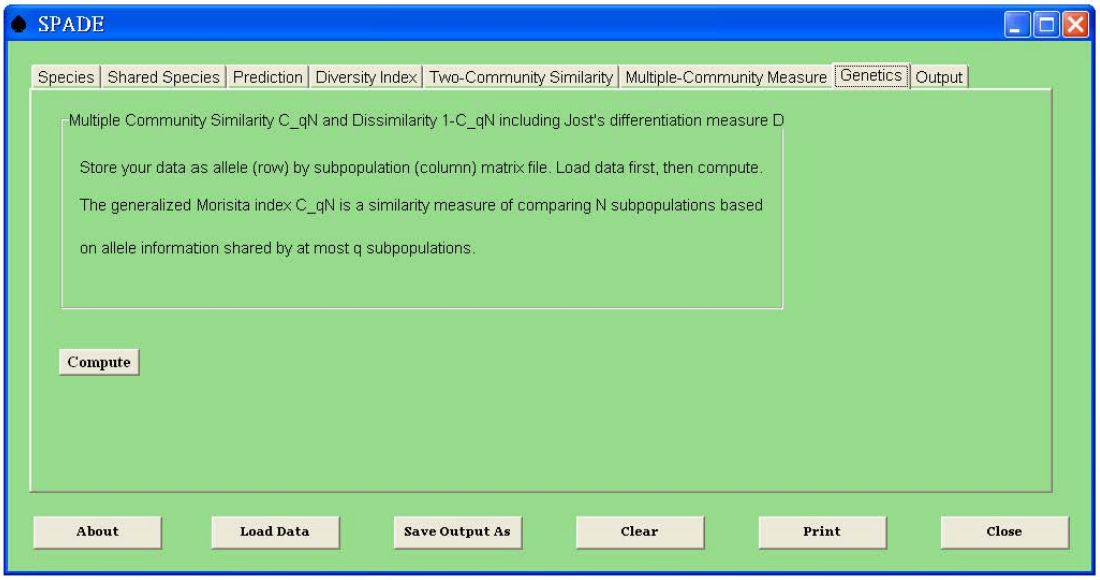
Part VII: Genetics (Estimating Allelic Differentiation/Similarity Among Subpopulations)

This part is almost identical to Part VI except that it is focused on genetic applications. The interface shown below is the same as that for Part VI except that "species" there is replaced by "alleles" and "community" by "subpopulation".

As in Part VI, store your data as an allele (row) by subpopulation (column) matrix file. Load data first and then compute. Please read Part VI for running procedures and other details. In many genetic data, only allele proportions are given. To use SPADE, you must convert proportional data (e.g. $p_j = 0.04$) to frequency data (e.g. 36, the actual number of sampled instances of allele j) first and then load data.

Remark: For estimating genetic diversity within each subpopulation, Part IV (Diversity Index) of SPADE provides the number of effective alleles (true diversity of orders 0, 1 and 2) and their s.e. estimate and 95% confidence interval. Refer to Part IV and example there for interpretation of output (replacing "species" and "community" there by "alleles" and "subpopulation"). See Jost (2006, 2007, 2008) for theoretical backgrounds.

Figure 7: The Interface Window of SPADE for Genetics Part



Example 7a: Genetics Data (Abundance/Frequency Data)

The hypothetical data in **Data7a.txt** include the frequencies from three subpopulations. The data are listed below. As described in Data Inputs Format for multiple communities, for any allele which is not found in the sample from a subpopulation, you must store frequency 0 in the data file. For example, in the following data, the first 13 alleles were not found in the second subpopulations. Thus we have 0's in the first 13 rows for the second column.

Table 7.1. Hypothetical Allele by Subpopulation Matrix

24	0	1
20	0	3

28	0	4
20	0	5
24	0	7
15	0	8
18	0	9
24	0	10
24	0	11
20	0	13
8	0	14
15	0	15
13	0	16
0	38	18
0	40	19
0	30	20
0	33	21
0	43	23
0	35	24
0	33	13

The output is shown below:

```
(1) BASIC DATA INFORMATION:

  The loaded set includes abundance (or frequency) data from 3 subpopulations
  and a total of 20 alleles.


  (Sample size in each subpopulation)                    n1=253
                                                         n2=252
                                                         n3=254

  (Number of alleles in one subpopulation)               D1=13
                                                         D2=7
                                                         D3=20

  (Number of shared alleles in two subpopulations)       D12=0
                                                         D13=13
                                                         D23=7

  (Bootstrap replications for s.e. estimate)             200

(2) NEARLY UNBIASED ESTIMATION OF ALLELIC DIFFERENTIATION OR MORISITA DISSIMILARITY IN 3 SUBPOPULATIONS

  Estimator   Estimate    Est_s.e.    95% Confidence Interval_1     95% Confidence Interval_2
    ------------------------------------------------------------------------------------------------------
  1-C23       0.608       0.010       (0.587, 0.628)                (0.590, 0.628)
  1-C23*      0.608       0.010       (0.587, 0.628)                (0.590, 0.628)
    ------------------------------------------------------------------------------------------------------
    *An adjusted estimator is recommended for practical use.

    1-C23: This is the genetic diversity measure D defined in Jost (2008) for comparing 20 subpopulations based
```

on allele shared information between any two subpopulations.

Confidence Interval_1: Based on estimate (+/-) 1.96 bootstrap s.e.
Confidence Interval_2: Based on an improved bootstrap percentile method. (recommend for use in the case when similarity is close to 0 or 1)

Pairwise Comparison:

1-C22(1,2)	1.000##	0.000##	(1.000, 1.000)##	(1.000, 1.000)##
1-C22(1,3)	0.540	0.048	(0.446, 0.634)	(0.451, 0.631)
1-C22(2,3)	0.225	0.037	(0.152, 0.298)	(0.162, 0.300)

Average Pairwise = 0.588

Pairwise Comparison (based on adjusted estimator):

1-C22(1,2)	1.000##	0.000##	(1.000, 1.000)##	(1.000, 1.000)##
1-C22(1,3)	0.540	0.048	(0.447, 0.634)	(0.451, 0.631)
1-C22(2,3)	0.225	0.037	(0.153, 0.298)	(0.163, 0.301)

Average Pairwise = 0.589

There are no shared species, thus estimated dissimilarity is one and should be used for caution.

1-C22: This is the genetic diversity measure D defined in Jost (2008) for comparing 2 subpopulations.

Dissimilarity Matrix

1-C22(i,j)	1	2	3
1	0.000	1.000	0.540
2		0.000	0.225
3			0.000

Dissimilarity Matrix(based on adjusted estimator)

1-C22(i,j)	1	2	3
1	0.000	1.000	0.540
2		0.000	0.225
3			0.000

Remark: If an estimator is less than 0, replace it by 0; if an estimator is greater than 1, replace it by 1.

(3) NEARLY UNBIASED ESTIMATION OF MORISITA SIMILARITY IN 3 SUBPOPULATIONS

Estimator	Estimate	Est_s.e.	95% Confidence Interval_1	95% Confidence Interval_2
C23	0.392	0.010	(0.372, 0.413)	(0.372, 0.410)
C23*	0.392	0.010	(0.372, 0.413)	(0.372, 0.410)

```
  *An adjusted estimator is recommended for practical use.

  C23: A similarity measure of comparing 3 subpopulations based on allele shared information between
       any two subpopulations.

  Pairwise Comparison:
  ------------------------------------------------------------------------------------------------------------
C22(1,2)    0.000##     0.000##     (0.000, 0.000)##            (0.000, 0.000)##
C22(1,3)    0.460       0.048       (0.366, 0.554)              (0.369, 0.549)
C22(2,3)    0.775       0.037       (0.702, 0.848)              (0.700, 0.838)
  ------------------------------------------------------------------------------------------------------------
  Average Pairwise =  0.412
  ------------------------------------------------------------------------------------------------------------

  Pairwise Comparison (based on adjusted estimator):
  ------------------------------------------------------------------------------------------------------------
C22(1,2)    0.000##     0.000##     (0.000, 0.000)##            (0.000, 0.000)##
C22(1,3)    0.460       0.048       (0.366, 0.553)              (0.369, 0.549)
C22(2,3)    0.775       0.037       (0.702, 0.847)              (0.699, 0.837)
  ------------------------------------------------------------------------------------------------------------
  Average Pairwise =  0.411
  ------------------------------------------------------------------------------------------------------------
  ## There are no shared species, thus estimated similarity is zero and should be used for caution.


  Similarity Matrix
  ------------------------------------------------------------------------------------------------------------
  C22(i,j)  1     2     3
1         1.000 0.000 0.460
2               1.000 0.775
3                     1.000
  ------------------------------------------------------------------------------------------------------------

  Similarity Matrix(based on adjusted estimator)
  ------------------------------------------------------------------------------------------------------------
  C22(i,j)  1     2     3
1         1.000 0.000 0.460
2               1.000 0.775
3                     1.000
  ------------------------------------------------------------------------------------------------------------
  Remark: If an estimator is less than 0, replace it by 0; if an estimator
   is greater than 1, replace it by 1.



References:
Chao, A., Jost, L., Chiang, S. C., Jiang, Y.-H. and Chazdon, R. (2008). A Two-
    stage probabilistic approach to multiple-community similarity indices.
    Biometrics, 64, 1178-1186.
Jost, L. (2008). GST and its relatives do not measure differentiation. Molecular
    Ecology, 17, 4015-4026.
```

To interpret the results, please see Example 6a. Here we specifically use a hypothetic example in which subpopulations 1 and 2 are completely differentiated (i.e., they have no shared alleles). A widely used genetic differentiation measure is Nei's G_{ST} based on the heterozygosity measure (Nei, 1973). We use this example to compare G_{ST} and Jost (2008) genetic differentiation measure D (D is identical to our dissimilarity measure $1-C_{22}$ for two-subpopulation case and $1-C_{23}$ for

three-subpopulation case). In Table 7.2, we show Nei's G_{ST} (not featured in SPADE) and Jost's D (given in our output) for three pairwise comparisons and one three-subpopulation comparison.

Table 7.2. Comparison of Two Measures: G_{ST} and Jost's D for Example 7a

subpopulations	Nei's G_{ST}	Jost's D	D in SPADE Output
(1, 2)	0.060	1.000	1− C22(1,2)
(1, 3)	0.022	0.540	1− C22(1,3)
(2, 3)	0.014	0.225	1− C22(2,3)
(1, 2, 3)	0.043	0.608	1− C23

Since the theoretical range for G_{ST} and for D is between 0 and 1, Jost's D correctly indicates that the two populations achieve complete differentiation whereas G_{ST} (0.06) indicates little differentiation. Also, Jost's D, with an estimate of 0.540 with 95% confidence interval of (0.446, 0.634), shows moderate differentiation between populations 1 and 3, but the value of G_{ST} (0.022) implies that there is almost no difference. Similar contrasting results are also shown in the three-subpopulation comparison. The estimate of D is 0.608 with a 95% confidence interval of (0.587, 0.628) based on ± 1.96 s.e. and (0.590, 0.628) based on a improved bootstrap percentile method, both signifying a relatively high difference. However, G_{ST} yields a value of 0.043 and thus implies a very low differentiation. See Jost (2006, 2007, 2008) for theoretical backgrounds and more examples of comparing these two measures.

There are two methods for constructing 95% confidence intervals:

(1) Confidence Interval_1: Based on estimate ± 1.96 bootstrap s.e.
When an estimating target parameter is not near the boundary of parameter space (i.e., 0 or 1), the asymptotic normality of an estimate is typically valid, so a symmetric 95% CI based on a normal critical point 1.96 can be applied.

(2) Confidence Interval_2: Based on an improved bootstrap percentile method.
When a parameter is near boundary 0 or 1, then the asymptotic normality is no longer valid, (because the central limit theory used for proving asymptotic normality requires a condition of the existence of a derivative around the true parameter), and the bootstrap distribution is generally skewed. In this case, we have an asymmetric confidence interval. Thus, this method is especially recommended for use in the case when the parameter is close to 0 or 1.) Here it is hard to define "how close is close" and may be data-dependent, but this method is always valid for all parameters. If the parameter is not near any boundary, the two methods give very close results. Therefore, we would recommend the use of this method.

Example 7b: Human Alleles Data (Abundance/Frequency Data)

We chose allele frequency counts from four human subpopulations (BiakaPyg, Palestin, Bedouin and Druze) from the data provided by Rosenberg et al. (2002). The allele frequency data in locus D3S2427 for four populations stored in **Data7b.txt** are shown below.

Table 7.3 Human Allele Frequency Data in Locus D3S2427
For Four Subpopulations

Allele	Subpopulations			
Code	BiakaPyg	Palestin	Bedouin	Druze

203	0	6	3	1
209	11	0	0	0
213	5	0	0	0
215	3	0	3	0
217	1	0	0	0
219	0	2	3	3
221	0	0	0	1
223	2	0	0	0
225	3	4	9	6
227	6	2	3	0
229	1	4	6	4
231	5	16	9	5
233	3	4	0	5
235	3	11	16	29
237	9	16	17	9
239	2	15	9	18
241	1	2	4	6
243	2	6	5	4
245	4	4	3	1
247	4	2	6	0
249	1	1	0	0
251	0	1	1	0
253	0	0	1	0
257	2	0	0	0
259	0	0	0	0
261	2	0	0	0
263	0	0	0	0

The output is as follows:

```
(1) BASIC DATA INFORMATION:

      The loaded set includes abundance (or frequency) data from 4 subpopulations
      and a total of 25 alleles.

      (Sample size in each subpopulation)          n1=70
                                                   n2=96
                                                   n3=98
                                                   n4=92

      (Number of alleles in one subpopulation)     D1=20
                                                   D2=16
                                                   D3=16
                                                   D4=13
```

(Number of shared alleles in two subpopulations) D12=13
D13=12
D14=10
D23=14
D24=12
D34=11

(Bootstrap replications for s.e. estimate) 200

(2) NEARLY UNBIASED ESTIMATION OF ALLELIC DIFFERENTIATION OR MORISITA DISSIMILARITY IN 4 SUBPOPULATIONS

Estimator	Estimate	Est_s.e.	95% Confidence Interval_1	95% Confidence Interval_2
1-C24	0.258	0.052	(0.157, 0.360)	(0.170, 0.362)
1-C24*	0.259	0.052	(0.157, 0.360)	(0.170, 0.362)

*An adjusted estimator is recommended for practical use.

1-C24: This is the genetic diversity measure D defined in Jost (2008) for comparing 20 subpopulations based on allele shared information between any two subpopulations.

Confidence Interval_1: Based on estimate (+/-) 1.96 bootstrap s.e.
Confidence Interval_2: Based on an improved bootstrap percentile method. (recommend for use in the case when similarity is close to 0 or 1)

Pairwise Comparison:

1-C22(1,2)	0.338	0.099	(0.145, 0.532)	(0.145, 0.531)
1-C22(1,3)	0.290	0.094	(0.105, 0.475)	(0.122, 0.454)
1-C22(1,4)	0.602	0.087	(0.431, 0.774)	(0.424, 0.766)
1-C22(2,3)	0.015	0.064	(0.000, 0.140)**	(0.000, 0.156)**
1-C22(2,4)	0.188	0.093	(0.007, 0.370)	(0.022, 0.385)
1-C22(3,4)	0.137	0.079	(0.000, 0.292)**	(0.003, 0.310)

Average Pairwise = 0.262

Pairwise Comparison (based on adjusted estimator):

1-C22(1,2)	0.341	0.099	(0.147, 0.534)	(0.147, 0.534)
1-C22(1,3)	0.293	0.094	(0.108, 0.478)	(0.125, 0.457)
1-C22(1,4)	0.604	0.087	(0.432, 0.775)	(0.425, 0.767)
1-C22(2,3)	0.017	0.064	(0.000, 0.142)**	(0.000, 0.158)**
1-C22(2,4)	0.190	0.093	(0.008, 0.371)	(0.023, 0.387)
1-C22(3,4)	0.139	0.079	(0.000, 0.294)**	(0.005, 0.312)

Average Pairwise = 0.264

**If the lower bound is less than 0, it is replaced by 0; if the upper bound is greater than 1, it is replaced by 1.

1-C22: This is the genetic diversity measure D defined in Jost (2008) for comparing 2 subpopulations.

Dissimilarity Matrix

1-C22(i,j)	1	2	3	4
1	0.000	0.338	0.290	0.602
2		0.000	0.015	0.188
3			0.000	0.137
4				0.000

Dissimilarity Matrix(based on adjusted estimator)

1-C22(i,j)	1	2	3	4
1	0.000	0.341	0.293	0.604
2		0.000	0.017	0.190

```
  3                    0.000 0.139
  4                          0.000
-------------------------------------------------------------------------------------------------------------
Remark: If an estimator is less than 0, replace it by 0; if an estimator
 is greater than 1, replace it by 1.
```

(3) NEARLY UNBIASED ESTIMATION OF MORISITA SIMILARITY IN 4 SUBPOPULATIONS

Estimator	Estimate	Est_s.e.	95% Confidence Interval_1	95% Confidence Interval_2
C24	0.742	0.052	(0.640, 0.843)	(0.638, 0.830)
C24*	0.741	0.052	(0.640, 0.843)	(0.638, 0.830)

*An adjusted estimator is recommended for practical use.

C24: A similarity measure of comparing 4 subpopulations based on allele shared information between any two subpopulations.

Pairwise Comparison:

C22(1,2)	0.662	0.099	(0.468, 0.855)	(0.469, 0.855)
C22(1,3)	0.710	0.094	(0.525, 0.895)	(0.546, 0.878)
C22(1,4)	0.398	0.087	(0.226, 0.569)	(0.234, 0.576)
C22(2,3)	0.985	0.064	(0.860, 1.000)**	(0.844, 1.000)
C22(2,4)	0.812	0.093	(0.630, 0.993)	(0.615, 0.978)
C22(3,4)	0.863	0.079	(0.708, 1.000)**	(0.690, 0.997)

Average Pairwise = 0.738

Pairwise Comparison (based on adjusted estimator):

C22(1,2)	0.659	0.099	(0.466, 0.853)	(0.466, 0.853)
C22(1,3)	0.707	0.094	(0.522, 0.892)	(0.543, 0.875)
C22(1,4)	0.396	0.087	(0.225, 0.568)	(0.233, 0.575)
C22(2,3)	0.983	0.064	(0.858, 1.000)**	(0.842, 1.000)**
C22(2,4)	0.810	0.093	(0.629, 0.992)	(0.613, 0.977)
C22(3,4)	0.861	0.079	(0.706, 1.000)**	(0.688, 0.995)

Average Pairwise = 0.736

**If the lower bound is less than 0, it is replaced by 0; if the upper bound is greater than 1, it is replaced by 1.

Similarity Matrix

C22(i,j)	1	2	3	4
1	1.000	0.662	0.710	0.398
2		1.000	0.985	0.812
3			1.000	0.863
4				1.000

Similarity Matrix(based on adjusted estimator)

C22(i,j)	1	2	3	4
1	1.000	0.659	0.707	0.396
2		1.000	0.983	0.810
3			1.000	0.861
4				1.000

Remark: If an estimator is less than 0, replace it by 0; if an estimator is greater than 1, replace it by 1.

References:
Chao, A., Jost, L., Chiang, S. C., Jiang, Y.-H. and Chazdon, R. (2008). A Two-stage probabilistic approach to multiple-community similarity indices. Biometrics, 64, 1178-1186.
Jost, L. (2008). GST and its relatives do not measure differentiation. Molecular Ecology, 17, 4015-4026.

Again, please see Example 6a for output interpretations. Here we remark that in the third part of the output, the lower bound of a 95% confidence interval may be below 0 due to sampling variation. In such cases, the lower bound is suggested to be replaced by the minimum value of 0. On the other hand, if an upper bound exceeds the theoretical maximum value of 1, it could be replaced by 1. In the following table, we compare Nei's G_{ST} and Jost's D for this real data set. All values of G_{ST} indicate little differentiations between any two subpopulations and among four subpopulations. In contrast, Jost'D shows that relatively high differentiation exists between populations 1 and 4. For overall comparison of the four populations, the estimate of Jost's D is 0.258 with a 95% confidence interval of (0.170, 0.362) based an improved bootstrap method, implying notable differentiation. However, the value of G_{ST} is 0.030 which shows almost no differentiation.

Table 7.4. Comparison of Two Measures: G_{ST} and Jost's D for Example 7b

subpopulations	Nei's G_{ST}	Jost's D	D in SPADE Output
(1, 2)	0.021	0.338	1− C22(1,2)
(1, 3)	0.018	0.290	1− C22(1,3)
(1, 4)	0.043	0.602	1− C22(1,4)
(2, 3)	0.006	0.015	1− C22(2,3)
(2, 4)	0.019	0.188	1− C22(2,4)
(3, 4)	0.015	0.137	1− C22(3,4)
(1, 2, 3, 4)	0.030	0.258	1− C24

Acknowledgements:

We sincerely thank a number of users for feedbacks and comments, which have led the removal of several bugs in SPADE. We acknowledge the National Science Council of Taiwan for continued support of our research projects related to SPADE.

References

Boneh, S., Boneh, A. and Caron, R. J. (1998). Estimating the prediction function and the number of unseen species in sampling with replacement. *Journal of the American Statistical Association* **93,** 372-379.

Bunge, J. and Fitzpatrick, M. (1993). Estimating the number of species: A review. *Journal of the American Statistical Association* **88**, 364-373.

Burnham, K. P. and Overton, W. S. (1978). Estimation of the size of a closed population when

capture probabilities vary among animals. *Biometrika* **65**, 625-633.

Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics* **11,** 265-270.

Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* **43**, 783-791.

Chao, A. (2005). Species estimation and applications. *Encyclopedia of Statistical Sciences,* Second Edition, Vol. 12, 7907-7916 (N. Balakrishnan, C. B. Read and B. Vidakovic, Editors), Wiley, New York.

Chao, A. and Bunge, J. (2002). Estimating the number of species in a stochastic abundance model. *Biometrics*, **58**, 531-539.

Chao, A., Chazdon, R. L., Colwell, R. K. and Shen, T.-J. (2005). A new statistical approach for assessing similarity of species composition with incidence and abundance data. *Ecology Letters* **8**, 148-159.

Chao, A., Chazdon, R. L., Colwell, R. K. and Shen, T.-J. (2006a). Abundance-based similarity indices and their estimation when there are unseen species in samples. *Biometrics* **62**, 361-371.

Chao, A., Hwang, W.-H., Chen, Y.-C. and Kuo. C.-Y. (2000). Estimating the number of shared species in two communities. *Statistica Sinica* **10,** 227-246.

Chao, A., Jost, L., Chiang, S. C., Jiang, Y.-H. and Chazdon, R. (2008). A Two-stage probabilistic approach to multiple-community similarity indices. *Biometrics* **64**, 1178-1186.

Chao, A. and Lee, S.-M. (1992). Estimating the number of classes via sample coverage. *Journal of the American Statistical Association* **87**, 210-217.

Chao, A. Ma, M-C. and Yang, M. C. K. (1993). Stopping rule and estimation for recapture debugging with unequal detection rates. *Biometrika* **80**, 193-201.

Chao, A. and Shen, T.-J. (2003). Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environmental and Ecological Statistics* **10***,* 429-443.

Chao, A. and Shen, T.-J. (2004). Non-parametric prediction in species sampling. *Journal of Agricultural, Biological, and Environmental Statistics* **9**, 253-269.

Chao, A., Shen, T.-J. and Hwang, W. H. (2006b). Application of Laplace's boundary-mode approximations to estimate species and shared species richness. *Australian and New Zealand Journal of Statistics* **48**, 117-128.

Colwell, R. K. (1997). EstimateS: Statistical estimation of species richness and shared species from samples. Version 5. User's Guide and Application published at http://viceroy.eeb.uconn.edu/estimates.

Colwell, R. K. and Coddington, J. A. (1994). Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society of London B – Biological Sciences* **345**, 101-118.

Efron, B. and Thisted, R. (1976). Estimating the number of unseen species: how many words did Shakespeare know? *Biometrika* **63,** 435-447.

Fisher, R. A., Corbet, A. S. and Williams, C. B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology* **12**, 42-58.

Holst, L. (1981). Some asymptotic results for incomplete multinomial or Poisson samples. *Scandinavian Journal of Statistics* **8**, 243-246.

Janzen, D. H. (1973a). Sweep samples of tropical foliage insects: description of study sites, with data on species abundances and size distributions. *Ecology* **54**, 659-686.

Janzen, D. H. (1973b). Sweep samples of tropical foliage insects: effects of seasons, vegetation types, elevation, time of day, and insularity. *Ecology* **54**, 687-708.

Jost, L. (2006). Entropy and diversity. *Oikos* **113**, 363-375.

Jost, L. (2007). Partitioning diversity into independent alpha and beta components. *Ecology* **88**, 2427-2439.

Jost, L. (2008). GST and its relatives do not measure differentiation. *Molecular Ecology* **17**, 4015-4026.

Krebs, C. J. (1999). *Ecological Methodology*, Second Edition, Harper & Row, New York.

Lee, S.-M. and Chao, A. (1994). Estimating population size via sample coverage for closed capture-recapture models. *Biometrics* **50**, 88-97.

Longino, J. T., Coddington, J. & Colwell, R. K. (2002). The ant fauna of a tropical rain forest: estimating species richness three different ways. *Ecology* **83**, 689-702.

Magurran, A. E. (1988). *Ecological Diversity and Its Measurement.* Princeton, Princeton University Press, New Jersey.

Magurran A. E. (2004). *Measuring Biological Diversity*. Blackwell, Oxford.

Miller, R. I. and Wiegert, R. G. (1989). Documenting completeness, species-area relations, and the species-abundance distribution of a regional flora. *Ecology* **70,** 16-22.

Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences, USA* **70**, 3321-3323.

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A. and Feldman, M. W. (2002). Genetic structure of human populations. *Science* **298**, 2381-2385.

Shen, T.- J. (2003). *Prediction of Biodiversity*, Ph.D. dissertation, National Tsing-Hua University, Hsin-Chu, Taiwan.

Shen, T.-J., Chao, A. and Lin, J.-F. (2002). Predicting the number of new species in further taxonomic sampling. *Ecology* **84**, 798-804.

Solow, A. R. and Polasky, S. (1999). A quick estimator for taxonomic surveys. *Ecology* **80,** 2799-2803.

Williams, C. B. (1964), *Patterns in the Balance of Nature*. London: Academic Press.

Zahl, S. (1977). Jackknifing an index of diversity. *Ecology* **58**, 907-913.

Appendix: Description and Formulas for Models/Estimators

Common Notation to All Parts:

S total number of species in a community.

X_i number of individuals (frequency) the ith species is observed in the sample, $i = 1, 2, \ldots, S$; (Only those species with $X_i > 0$ are observable in the sample).

n sample size, $n = \sum_{i=1}^{S} X_i = \sum_{j \geq 1} j f_j$. (f's are defined below)

$I[A]$ the usual indicator function, i.e., $I[A] = 1$ if the event A occurs, 0 otherwise.

f_j number of species that are represented exactly j times in the sample, $j = 0, 1, \ldots, n$, $f_j = \sum_{i=1}^{S} I[X_i = j]$. (f_0 denotes the number of unobserved species).

C sample coverage

$\hat{}$ an estimator from data; e.g., $\hat{S}$ (estimator of S) and $\hat{C}$ (estimator of C)…and so on.

D number of distinct species discovered in the sample, ($D = \sum_{i=1}^{S} I[X_i > 0] = \sum_{j \geq 1} f_j$).

t number of samples/occasions/quadrats (for multiple incidence data).

Q_k number of species that are observed in exactly k samples, $k = 0, 1, \ldots, t$, based on presence/absence data.

κ the cut-off point (default = 10), which separates species into "abundant" and "rare" groups for abundance data; it separates species into "frequent" and "infrequent" groups for incidence data.

CV: coefficient of variation

Part I: Species (Species Richness Estimation in One Community)

Estimators in Examples 1a, 1b and 1d

Additional Notation for these examples:

n_{rare} : sample size for "rare" group.

D_{rare} : the number of distinct species for "rare" group; $D_{rare} = \sum_{i=1}^{\kappa} f_i$.

D_{abun} : the number of distinct species for "abundant" group; $D_{abun} = \sum_{i > \kappa} f_i$.

$\hat{C}_{rare}$: estimated sample coverage for "rare" group; $\hat{C}_{rare} = 1 - f_1 / \sum_{i=1}^{\kappa} i f_i$.

$\hat{\gamma}_{rare}$ or $\hat{CV}_{rare}$: estimated CV

$\tilde{\gamma}_{rare}$: estimated CV for highly heterogeneous case.

- Homogenous Model (Chao and Lee, 1992):

$$\hat{S} = D_{abun} + \frac{D_{rare}}{\hat{C}_{rare}},$$

where $D_{rare} = \sum_{i=1}^{\kappa} f_i$, $\hat{C}_{rare} = 1 - f_1 / \sum_{i=1}^{\kappa} i f_i$ and κ = cut-off point.

- Homogeneous (MLE): Approximate MLE under the homogeneous model

$\hat{S}$ is the solution of $D = S[1 - \exp(-n/S)]$.

- Chao1 (Chao, 1984):

$$\hat{S} = \begin{cases} D + f_1^2/(2f_2), & \text{if } f_2 > 0 \\ D + f_1(f_1 - 1)/2, & \text{if } f_2 = 0 \end{cases}$$

- Chao1-bc (bias-corrected estimator for Chao1)

$$\hat{S} = D + f_1(f_1 - 1)/[2(f_2 + 1)].$$

- ACE (Chao and Lee, 1992):

$$\hat{S} = D_{abun} + \frac{D_{rare}}{\hat{C}_{rare}} + \frac{f_1}{\hat{C}_{rare}} \hat{\gamma}^2_{rare},$$

where $D_{abun} = \sum_{i>\kappa} f_i$ and $\hat{\gamma}^2_{rare} = \max\left\{ \frac{D_{rare}}{\hat{C}_{rare}} \frac{\sum_{i=1}^{\kappa} i(i-1) f_i}{(\sum_{i=1}^{\kappa} i f_i)(\sum_{i=1}^{\kappa} i f_i - 1)} - 1, 0 \right\}$.

(CV_rare in the output)

- ACE-1 (Chao and Lee, 1992) for highly-heterogeneous case

$$\hat{S} = D_{abun} + \frac{D_{rare}}{\hat{C}_{rare}} + \frac{f_1}{\hat{C}_{rare}} \tilde{\gamma}^2_{rare}$$

where $\tilde{\gamma}^2_{rare} = max\left\{ \hat{\gamma}^2_{rare}\left(1 + \frac{(1 - \hat{C}_{rare}) \sum_{i=1}^{\kappa} i(i-1) f_i}{\hat{C}_{rare}(\sum_{i=1}^{\kappa} i f_i - 1)}\right), \quad 0 \right\}$. (CV1_rare in the output)

- 1st order jackknife (Abundance Data) (Burnham and Overton, 1978):

$$\hat{S} = D + \frac{n-1}{n} f_1.$$

- 2nd order jackknife (Abundance Data) (Burnham and Overton, 1978):

$$\hat{S} = D + \frac{2n-3}{n} f_1 - \frac{(n-2)^2}{n(n-1)} f_2.$$

- Gamma-Poisson Model (Chao and Bunge, 2002)

Assume that species are discovered in the sample according to a Poisson process and rates of the processes follow a gamma distribution with parameters (α, β).

$$\hat{S} = D_{abun} + \sum_{i=2}^{\kappa} f_i \Bigg/ \left(1 - \frac{f_1 \sum_{i=1}^{\kappa} i^2 f_i}{(\sum_{i=1}^{\kappa} i f_i)^2}\right).$$

- Gamma-Poisson-UMLE (Chao and Bunge, 2002):

  Under the Gamma-Poisson model, The UMLE of (S, α, β) are solutions of the following equations:

$$\frac{1}{S} + \frac{1}{S-1} + \ldots + \frac{1}{S-D+1} + \alpha \log(\frac{\beta}{\beta+T}) = 0,$$

$$S \log(\frac{\beta}{\beta+T}) + \frac{D}{\alpha} + \frac{D-f_1}{\alpha+1} + \frac{D-(f_1+f_2)}{\alpha+2} + \ldots + \frac{f_n}{\alpha+n-1} = 0,$$

$$\frac{\alpha S}{\beta} - \frac{\alpha S + n}{\beta + T} = 0.$$

- Gamma-Poisson-CMLE (Chao and Bunge, 2002):

  Under the Gamma-Poisson model, the CMLE of (S, α, β) are solutions of the following equations:

$$\frac{S-D}{\left(\frac{\beta}{\beta+T}\right)^{\alpha}} - \frac{D}{1-\left(\frac{\beta}{\beta+T}\right)^{\alpha}} = 0,$$

$$S \log(\frac{\beta}{\beta+T}) + \frac{D}{\alpha} + \frac{D-f_1}{\alpha+1} + \frac{D-(f_1+f_2)}{\alpha+2} + \ldots + \frac{f_n}{\alpha+n-1} = 0,$$

$$\frac{\alpha S}{\beta} - \frac{\alpha S + n}{\beta + T} = 0.$$

Estimators in Example 1c and 1e:

Additional Notation

D_{freq}: the number of species for "frequent" group; $D_{freq} = \sum_{i>\kappa} Q_i$.

D_{infreq}: the number of species for "infrequent" group; $D_{infreq} = \sum_{i=1}^{\kappa} Q_i$.

$\hat{C}_{infreq}$: estimated sample coverage for "infrequent" group; $\hat{C}_{infreq} = 1 - \frac{Q_1}{\sum_{i=1}^{\kappa} i Q_i} \frac{(t-1)Q_1}{[(t-1)Q_1 + 2Q_2]}$.

$\hat{\gamma}_{infreq}$ or $\hat{CV}_{infreq}$: estimated squared CV.

$\tilde{\gamma}_{infreq}$: estimated CV for highly heterogeneous case.

n_i: number of species in the ith sample for the "infrequent" group

- Homogeneous Model (Lee and Chao, 1994):

$$\hat{S} = D_{freq} + \frac{D_{infreq}}{\hat{C}_{infreq}}$$

where $D_{freq} = \sum_{i>\kappa} Q_i$, $D_{infreq} = \sum_{i=1}^{\kappa} Q_i$, and $\hat{C}_{infreq} = 1 - \frac{Q_1}{\sum_{i=1}^{\kappa} iQ_i} \frac{(t-1)Q_1}{[(t-1)Q_1 + 2Q_2]}$.

- Chao2 (Chao, 1987):

$$\hat{S} = \begin{cases} D + (t-1)Q_1^2/(2tQ_2), & \text{if } Q_2 > 0 \\ D + (t-1)Q_1(Q_1-1)/(2t), & \text{otherwise} \end{cases}$$

- Chao2-bc (bias-corrected form for Chao2)

$$\hat{S} = D + [(t-1)/t]Q_1(Q_1-1)/[2(Q_2+1)].$$

- Model (h) (ICE) (Lee and Chao, 1994):

$$\hat{S} = D_{freq} + \frac{D_{infreq}}{\hat{C}_{infreq}} + \frac{Q_1}{\hat{C}_{infreq}} \hat{\gamma}^2_{infreq},$$

where $\hat{\gamma}^2_{infreq} = max\left\{ \frac{D_{infreq}}{\hat{C}_{infreq}} \frac{t}{(t-1)} \frac{\sum_{i=1}^{\kappa} i(i-1)Q_i}{(\sum_{i=1}^{\kappa} iQ_i)(\sum_{i=1}^{\kappa} iQ_i - 1)} - 1,\ 0 \right\}$.

- Model(h)-1 (ICE-1) for highly-heterogeneous case:

$$\hat{S} = D_{freq} + \frac{D_{infreq}}{\hat{C}_{infreq}} + \frac{Q_1}{\hat{C}_{infreq}} \tilde{\gamma}^2_{infreq},$$

where $\tilde{\gamma}^2_{infreq} = max\left\{ (\hat{S}_{ICE}) \frac{t}{(t-1)} \frac{\sum_{i=1}^{\kappa} i(i-1)Q_i}{(\sum_{i=1}^{\kappa} iQ_i)(\sum_{i=1}^{\kappa} iQ_i - 1)} - 1,\ \ 0 \right\}$ and $\hat{S}_{ICE}$ denotes the estimate from Model(h) (ICE).

- Model (th) (Lee and Chao, 1994):

$$\hat{S} = D_{freq} + \frac{D_{infreq}}{\hat{C}_{infreq}} + \frac{Q_1}{\hat{C}_{infreq}} \hat{\gamma}^2_{infreq},$$

where $\hat{\gamma}^2_{infreq} = max\left\{ \frac{D_{infreq}}{\hat{C}_{infreq}} \frac{\sum_{i=1}^{\kappa} i(i-1)Q_i}{2\sum\sum_{j<k} n_j n_k} - 1,\ 0 \right\}$.

- Model (th)-1 (Lee and Chao, 1994):

$$\hat{S} = D_{freq} + \frac{D_{infreq}}{\hat{C}_{infreq}} + \frac{Q_1}{\hat{C}_{infreq}} \tilde{\gamma}^2_{infreq},$$

where $\tilde{\gamma}^2_{infreq} = max\left\{\hat{S}_{Model(th)} \frac{\sum_{i=1}^{\kappa} i(i-1)Q_i}{2\sum\sum_{j<k} n_j n_k} - 1,\ 0\right\}$ and $\hat{S}_{Model(th)}$ denotes the estimate from Model(th).

- 1st order jackknife (Incidence Data) (Burnham and Overton, 1978):

$$\hat{S} = D + \frac{t-1}{t} Q_1 .$$

- 2nd order jackknife (Incidence Data) (Burnham and Overton, 1978):

$$\hat{S} = D + \frac{2t-3}{t} Q_1 - \frac{(t-2)^2}{t(t-1)} Q_2 .$$

- Beta-binomial-CMLE

  The estimator is the conditional MLE for beta-binomial model (the number of samples that a species is detected follows a binomial distribution and the detection probability follows a beta distribution.)

- Beta-binomial-UMLE

  The estimator is the unconditional MLE for beta-binomial model (the number of samples that a species is detected follows a binomial distribution and the detection probability follows a beta distribution.)

Part II: Shared Species (Estimating Shared Spceies Richness in Two Communities)

Formulas for Example 2a:

Additional Notation

S_{12}: the number of shared species.

D_{12}: the number of shared species.

(X_i, Y_i): sample frequencies of the ith species in the two communities.

$D_{12(rare)}$: the number of shared species in the "rare" group,

$$D_{12(rare)} = \sum_{i=1}^{S_{12}} I(0 < X_i \le 10,\ 0 < Y_i \le 10) .$$

$D_{12(abun)}$: the number of shared species in the "abundant" group, $D_{12(abun)} = D_{12} - D_{12(rare)}$.

$f_{1+(rare)}$: number of shared species that are singletons in Community I for the "rare" group.

$$f_{1+(rare)} = \sum_{i=1}^{D_{12}} I(X_i = 1,\ 0 < Y_i \le 10) .$$

$f_{+1(rare)}$: number of shared species that are singletons in Community II for the "rare" group.

$$f_{+1(rare)} = \sum_{i=1}^{D_{12}} I(0 < X_i \le 10,\ Y_i = 1),$$

f_{11}: number of shared species that are singletons in both communities.

$$f_{11} = \sum_{i=1}^{D_{12}} I(X_i = Y_i = 1).$$

- Homogeneous (Chao et al., 2000):

$$\hat{S}_{12} = D_{12(abun)} + \frac{D_{12(rare)}}{\hat{C}_{12(rare)}},$$

where $D_{12(rare)} = \sum_{i=1}^{D_{12}} I(0 < X_i \le 10,\ 0 < Y_i \le 10)$ and

$$\hat{C}_{12(rare)} = 1 - \frac{\sum_{i=1}^{D_{12}} [Y_i I(X_i = 1,\ Y_i \le 10) + X_i I(Y_i = 1,\ X_i \le 10) - I(X_i = Y_i = 1)]}{\sum_{i=1}^{D_{12}} X_i Y_i I(X_i \le 10,\ Y_i \le 10)}.$$

- Heterogeneous (ACE-shared); see Chao et al. (2000):

$$\hat{S}_{12} = D_{12(abun)} + \frac{D_{12(rare)}}{\hat{C}_{12(rare)}} + \frac{1}{\hat{C}_{12(rare)}} \left[f_{1+(rare)} \hat{\Gamma}_1 + f_{+1(rare)} \hat{\Gamma}_2 + f_{11} \hat{\Gamma}_{12} \right],$$

where $\hat{S}_{12}^{(0)} = \dfrac{D_{12(rare)}}{\hat{C}_{12(rare)}}$

(CCV$_1$) $\hat{\Gamma}_1 = \dfrac{\hat{S}_{12}^0 n_{1(rare)} T_{21}}{(n_{1(rare)} - 1) T_{10} T_{11}} - 1,$

(CCV$_2$) $\hat{\Gamma}_2 = \dfrac{\hat{S}_{12}^0 n_{2(rare)} T_{12}}{(n_{2(rare)} - 1) T_{01} T_{11}} - 1,$

(CCV$_{12}$) $\hat{\Gamma}_{12} = \dfrac{n_{1(rare)} n_{2(rare)} (\hat{S}_{12}^0)^2 T_{22}}{(n_{1(rare)} - 1)(n_{2(rare)} - 1) T_{10} T_{01} T_{11}} - \dfrac{\hat{S}_{12}^0 T_{11}}{T_{10} T_{01}} - \hat{\Gamma}_1 - \hat{\Gamma}_2,$

$T_{10} = \sum_{i=1}^{D_{12}} X_i I(X_i \le 10, Y_i \le 10)$, $T_{01} = \sum_{i=1}^{D_{12}} Y_i I(X_i \le 10, Y_i \le 10)$,

$T_{11} = \sum_{i=1}^{D_{12}} X_i Y_i I(X_i \le 10, Y_i \le 10)$, $T_{21} = \sum_{i=1}^{D_{12}} X_i (X_i - 1) Y_i I(X_i \le 10, Y_i \le 10)$,

$T_{12} = \sum_{i=1}^{D_{12}} X_i Y_i (Y_i - 1) I(X_i \le 10, Y_i \le 10)$,

$T_{22} = \sum_{i=1}^{D_{12}} X_i (X_i - 1) Y_i (Y_i - 1) I(X_i \le 10, Y_i \le 10)$.

Notation for Chao1-shared and Chao1-shared-bc:

$f_{1+} = \sum_{i=1}^{D_{12}} I[X_i = 1, Y_i \ge 1]$: the observed number of shared species that occur once in Sample;

$f_{2+} = \sum_{i=1}^{D_{12}} I[X_i = 2, Y_i \geq 1]$: the observed number of shared species that occur twice in Sample 1;

$f_{+1} = \sum_{i=1}^{D_{12}} I[X_i \geq 1, Y_i = 1]$: the observed number of shared species that occur once in Sample 2;

$f_{+2} = \sum_{i=1}^{D_{12}} I[X_i \geq 1, Y_i = 2]$: the observed number of shared species that occur twice in Sample 2.

- Chao1-shared; see Chao, Shen and Hwang (2006):

$$\hat{S}_{12} = D_{12} + f_{11}\frac{f_{1+}f_{+1}}{4f_{2+}f_{+2}} + \frac{f_{1+}^2}{2f_{2+}} + \frac{f_{+1}^2}{2f_{+2}}.$$

- Chao1-shared-bc: (bias-corrected version) for f_{2+} = 0 or f_{+2} = 0

$$\hat{S}_{12}^* = D_{12} + f_{11}\frac{f_{1+}f_{+1}}{4(f_{2+}+1)(f_{+2}+1)} + \frac{f_{1+}(f_{1+}-1)}{2(f_{2+}+1)} + \frac{f_{+1}(f_{+1}-1)}{2(f_{+2}+1)}.$$

Formulas and Additional Notation for Example 2b:

Assume that there are t_1 samples in community I and there are t_2 samples in community II. Let X_i and Y_i denote the number of samples that the ith species are detected in communities I and II, respectively. Let $Q_{11} = \sum_{i=1}^{S_{12}} I(X_i = 1, Y_i = 1)$ denote the number of shared species that are detected only in one sample in both sets of samples. Similarly, define Q_{j+} and Q_{+k} as in the abundance case.

- Chao2-shared; see Chao, Shen and Hwang (2006):

$$\hat{S}_{12} = D_{12} + Q_{11}\frac{(t_1-1)}{t_1}\frac{(t_2-1)}{t_2}\frac{Q_{1+}Q_{+1}}{4Q_{2+}Q_{+2}} + \frac{(t_1-1)}{t_1}\frac{Q_{1+}^2}{2Q_{2+}} + \frac{(t_2-1)}{t_2}\frac{Q_{+1}^2}{2Q_{+2}}.$$

- Chao2-shared-bc: (bias-corrected version) for Q_{2+} = 0 or Q_{+2} = 0

$$\hat{S}_{12}^* = D_{12} + Q_{11}\frac{(t_1-1)(t_2-1)}{t_1t_2}\frac{Q_{1+}Q_{+1}}{4(Q_{2+}+1)(Q_{+2}+1)} + \frac{(t_1-1)}{t_1}\frac{Q_{1+}(Q_{1+}-1)}{2(Q_{2+}+1)} + \frac{(t_2-1)}{t_2}\frac{Q_{+1}(Q_{+1}-1)}{2(Q_{+2}+1)}.$$

Part III: Prediction (Predicting the Number of New Spceies in Further Survey)

Formulas for Example 3a (Multinomial Model):

Additional Notation:

m: prediction size

$S(m)$: the expected number of new species that will be observed in a prediction sample of size m.

- Efron and Thisted (1976)

$$\hat{S}(m) = \sum_{i=1}^{n} (-1)^{i+1} (\frac{m}{n})^{i} f_i ,$$

- Boneh, Boneh and Caron (1998)

$$\hat{S}(m) = \sum_{i=1}^{S} (1 - \frac{X_i}{n})^{n} [1 - (1 - \frac{X_i}{n})^{m}] + \nu (1 - \frac{f_1}{n\nu})^{n} [1 - (1 - \frac{f_1}{n\nu})^{m}] ,$$

where ν is the solution of the equation

$\nu\{1 - [1 - f_1 /(n\nu)]^n\} = \sum_{i=1}^{n} f_i (1 - i/n)^n$ provided the condition

$f_1 > \sum_{i=1}^{n} f_i \exp(-i)$ is satisfied.

- Solow and Polasky (1999):

$$\hat{S}(m) = \frac{f_1^2}{2 f_2} [1 - (1 - \frac{2 f_2}{n f_1})^{m}] .$$

- Shen, Chao and Lin (2003):

$\hat{S}(m) = \hat{f}_0 [1 - (1 - \frac{f_1}{n\hat{f}_0})^{m}]$, where $\hat{f}_0 = \frac{D_{rare}}{\hat{C}_{rare}} + \frac{f_1}{\hat{C}_{rare}} \hat{\gamma}^2_{rare} - D_{rare}$. (See Part I).

Formulas for Example 3b (Poisson Model):

Additional Notation:

T: prediction time interval.

S(*T*): the expected number of new species that will be observed in a prediction interval of length *T*.

- Efron and Thisted (1976):

$$\hat{S}(T) = \sum_{i=1}^{n} (-1)^{i+1} T^i f_i$$

- Boneh, Boneh and Caron (1998):

$$\hat{S}(T) = \sum_{k=1}^{n} f_k \exp(-k)[1 - \exp(-kT)] + \nu\{\exp(-f_1/\nu) - \exp[-f_1(1+T)/\nu] ,$$

where ν is the solution of the equation $\nu[-\exp(-f_1/\nu)] = \sum_{i=1}^{n} f_i \exp(-i)$

provided the condition $f_1 > \sum_{i=1}^{n} f_i \exp(-i)$ is satisfied.

- Chao and Shen (2004):

$\hat{S}(T) = \hat{f}_0 [1 - \exp(-T f_1 / \hat{f}_0)]$, where $\hat{f}_0 = \frac{D_{rare}}{\hat{C}_{rare}} + \frac{f_1}{\hat{C}_{rare}} \hat{\gamma}^2_{rare} - D_{rare}$ (See Part I) .

Part IV: Diversity Index (Estimating Various Diversity Indices in One Community)

Formulas for Examples 4a and Example 4b:

Shannon's Index $\hat{H}$ and its effective number of species, $\exp(\hat{H})$ (true diversity of order 1):

- MLE

$$\hat{H} = -\sum_{i=1}^{S} I(X_i > 0)\frac{X_i}{n}\log(\frac{X_i}{n}).$$

- MLE_bc:

$$\hat{H} = -\sum_{i=1}^{S} I(X_i > 0)\frac{X_i}{n}\log(\frac{X_i}{n}) + \frac{\hat{S}-1}{2n},$$

where $\hat{S} = D_{abun} + \frac{D_{rare}}{\hat{C}_{rare}} + \frac{f_1}{\hat{C}_{rare}}\hat{\gamma}^2_{rare}$ (ACE, see Part 1 and Example 1a).

- Jackknife (Zahl, 1977):

$$\hat{H} = n\log(n) - (n-1)\log(n-1) + \frac{1}{n}\sum_{k=2}^{n} f_k k^2 \log(\frac{k-1}{k}) - \frac{1}{n}\sum_{k=2}^{n} f_k k \log(k-1).$$

- Chao and Shen (2003):

$$\hat{H} = -\sum_{k=1}^{n} f_k \frac{(k(1-f_1/n)/n)\log[k(1-f_1/n)/n]}{1-[1-k(1-f_1/n)/n]^n}.$$

Simpson Index $\hat{\theta}$ and its effective number of species $1/\hat{\theta}$ (true diversity of order 2):

- MVUE:

$$\hat{\theta} = \sum_{k=1}^{n} f_k \frac{k(k-1)}{n(n-1)}.$$

- MLE:

$$\hat{\theta} = \sum_{k=1}^{n} f_k (\frac{k}{n})^2.$$

Fisher's Alpha Index (Fisher et al., 1943):

$$\hat{\alpha} = \frac{n(1-\hat{\phi})}{\hat{\phi}} \text{ where } \phi \text{ is solved from } \frac{D}{n} = \frac{1-\phi}{\phi}[-\ln(1-\phi)].$$

Part V: Two-Community Similarity

Formulas for Example 5a (Abundance Data):

D_i: the number of observed species in the Sample i, $i = 1,2$.

D_{12}: the number of observed shared species in the two samples.

(X_i, Y_i): sample frequencies of the ith species in the two samples.

$f_{1+} = \sum_{i=1}^{D_{12}} I[X_i = 1, Y_i \geq 1]$: the observed number of shared species that occur once in Sample 1 (these species must be present in Sample 2, but may have any frequency).

$f_{2+} = \sum_{i=1}^{D_{12}} I[X_i = 2, Y_i \geq 1]$: the observed number of shared species that occur twice in Sample 1.

$f_{+1} = \sum_{i=1}^{D_{12}} I[X_i \geq 1, Y_i = 1]$: the observed number of shared species that occur once in Sample 2 (these species must be present in Sample 1, but may have any frequency).

$f_{+2} = \sum_{i=1}^{D_{12}} I[X_i \geq 1, Y_i = 2]$: the observed number of shared species that occur twice in Sample 2.

$\widetilde{U} = \sum_{i=1}^{D_{12}} X_i / n$, $\widetilde{V} = \sum_{i=1}^{D_{12}} Y_i / m$,

$$\hat{U} = \sum_{i=1}^{D_{12}} \frac{X_i}{n} + \frac{(m-1)}{m} \frac{f_{+1}}{2f_{+2}} \sum_{i=1}^{D_{12}} \frac{X_i}{n} I(Y_i = 1),$$

$$\hat{V} = \sum_{i=1}^{D_{12}} \frac{Y_i}{m} + \frac{(n-1)}{n} \frac{f_{1+}}{2f_{2+}} \sum_{i=1}^{D_{12}} \frac{Y_i}{m} I(X_i = 1).$$

- Jaccard incidence:

$$\frac{D_{12}}{D_1 + D_2 - D_{12}}$$

- Sorensen incidence:

$$\frac{2D_{12}}{D_1 + D_2}$$

- Lennon et al (2001):

$$\frac{D_{12}}{D_{12} + \min(D_1 - D_{12}, D_2 - D_{12})}$$

- Bray-Curtis:

$$\hat{C}_{BC} = \frac{2\sum_{i=1}^{D_{12}} \min(X_i, Y_i)}{\sum_{i=1}^{D_1} X_i + \sum_{i=1}^{D_2} Y_i}$$

- Morisita-Horn:

$$\hat{C}_{MH} = \frac{2\sum_{i=1}^{D_{12}} \frac{X_i}{n} \frac{Y_i}{m}}{\sum_{i=1}^{D_1} \left(\frac{X_i}{n}\right)^2 + \sum_{i=1}^{D_2} \left(\frac{Y_i}{m}\right)^2}$$

- Morisita Original:

$$\hat{C}_{OM} = \frac{2\sum_{i=1}^{D_{12}} \frac{X_i}{n}\frac{Y_i}{m}}{\sum_{i=1}^{D_1}\left(\frac{X_i}{n}\right)\left(\frac{X_i-1}{n-1}\right)+\sum_{i=1}^{D_2}\left(\frac{Y_i}{m}\right)\left(\frac{Y_i-1}{m-1}\right)}$$

- Jaccard Abundance (unadjusted):

$$\frac{\tilde{U}\tilde{V}}{\tilde{U}+\tilde{V}-\tilde{U}\tilde{V}}$$

- Jaccard Abundance (adjusted):

$$\frac{\hat{U}\hat{V}}{\hat{U}+\hat{V}-\hat{U}\hat{V}}$$

- Sorensen Abundance (unadjusted):

$$\frac{2\tilde{U}\tilde{V}}{\tilde{U}+\tilde{V}}$$

- Sorensen Abundance (adjusted):

$$\frac{2\hat{U}\hat{V}}{\hat{U}+\hat{V}}$$

Formulas for Example 5b (Incidence Data):

For multiple incidence data: Suppose we take a set of w replicated incidence samples from Community 1 and a set of z replicated incidence samples from Community 2. Define (X_i, Y_i) as the sample incidence frequencies of the ith species in the two sets of multiple incidence records. (That is, X_i denotes the number of samples that the ith species is present in Sample 1, and similarly for Y_i for Sample 2.) All formulas are similar to those for abundance data, except that

(a) replace sample size n and m in the above by $n = \sum_{i=1}^{S} X_i$ and $m = \sum_{i=1}^{S} Y_i$;

(b) replace $\hat{U}$ and $\hat{V}$ respectively by

$$\hat{U} = \sum_{i=1}^{D_{12}} \frac{X_i}{n} + \frac{(w-1)}{w}\frac{f_{+1}}{2f_{+2}}\sum_{i=1}^{D_{12}}\left[\frac{X_i}{n} I(Y_i = 1)\right],$$

$$\hat{V} = \sum_{i=1}^{D_{12}} \frac{Y_i}{m} + \frac{(z-1)}{z}\frac{f_{1+}}{2f_{2+}}\sum_{i=1}^{D_{12}}\left[\frac{Y_i}{m} I(X_i = 1)\right].$$

Part VI: Multiple-Community Measure

Formulas for Example 6a

Notation: Assume that there are S species in the combined assemblage and let the species be indexed by 1, 2, …, S. Define N sets of probabilities of species discovery (i.e., relative abundance)

as $\{(p_{1r}, p_{2r},..., p_{Sr}); r = 1, ..., N\}$, where $p_{ir} \geq 0$. Assume that a random sample of n_j individuals is taken from Community j for $j = 1, 2, \ldots, N$. Denote the N sets of sample frequencies by $\{(X_{1j}, X_{2j},..., X_{Sj}); j = 1, 2, ..., N\}$.

- $N = 2$ (Two Communities)

  The widely used Morisita abundance-based index (Krebs, 1999, p. 390) is

$$C_{22} = \frac{2\sum_{i=1}^{S} p_{i1}p_{i2}}{\sum_{i=1}^{S} p_{i1}^2 + \sum_{i=1}^{S} p_{i2}^2}.$$

  A nearly unbiased estimator of C_{22} is identical to the original Morisita index in Part V and has the following form:

$$\hat{C}_{22} = \frac{2\sum_{i=1}^{S} \frac{X_{i1}}{n_1}\frac{X_{i2}}{n_2}}{\sum_{i=1}^{S} \frac{X_{i1}(X_{i1}-1)}{n_1(n_1-1)} + \sum_{i=1}^{S} \frac{X_{i2}(X_{i2}-1)}{n_2(n_2-1)}}$$

  However, in some cases, this estimator may exceed the theoretical maximum value of 1. In such cases, it is replaced by its MLE (i.e., the Morisita-Horn index in Part V):

$$\tilde{C}_{22} = \frac{2\sum_{i=1}^{S} \frac{X_{i1}}{n_1}\frac{X_{i2}}{n_2}}{\sum_{i=1}^{S} \left(\frac{X_{i1}}{n_1}\right)^2 + \sum_{i=1}^{S} \left(\frac{X_{i2}}{n_2}\right)^2}.$$

- $N = 3$ (Three Communities)

  The nearly unbiased estimate for C_{23} (C_{23} formula is given in the general N communities cases) is

$$\hat{C}_{23} = \frac{\sum_{i=1}^{S}\left[\frac{X_{i1}X_{i2}}{n_1n_2} + \frac{X_{i1}X_{i3}}{n_1n_3} + \frac{X_{i2}X_{i3}}{n_2n_3}\right]}{\sum_{i=1}^{S}\left[\frac{X_{i1}^{(2)}}{n_1^{(2)}} + \frac{X_{i2}^{(2)}}{n_2^{(2)}} + \frac{X_{i3}^{(2)}}{n_3^{(2)}}\right]},$$

  where $x^{(k)} = x(x-1)...(x-k+1)$. The nearly unbiased estimate for C_{33} (C_{33} formula is given in the general N communities cases) is:

$$\hat{C}_{33} = \frac{\frac{1}{24}\sum_{i=1}^{S}\left[3\frac{X_{i1}^{(2)}X_{i2}}{n_1^{(2)}n_2} + 3\frac{X_{i1}X_{i2}^{(2)}}{n_1 n_2^{(2)}} + 3\frac{X_{i1}^{(2)}X_{i3}}{n_1^{(2)}n_3} + ... + 6\frac{X_{i1}X_{i2}X_{i3}}{n_1 n_2 n_3}\right]}{\frac{1}{3}\sum_{i=1}^{S}\left[\frac{X_{i1}^{(3)}}{n_1^{(3)}} + \frac{X_{i2}^{(3)}}{n_2^{(3)}} + \frac{X_{i3}^{(3)}}{n_3^{(3)}}\right]}.$$

- General *N* Communities

The theoretical formula for C_{qN} is

$$C_{qN} = \frac{\frac{1}{(N^q - N)}\sum_{i=1}^{S}\left[(p_{i1} + p_{i2} + ... + p_{iN})^q - (p_{i1}^q + p_{i2}^q + ...p_{iN}^q)\right]}{\frac{1}{N}\sum(p_{i1}^q + p_{i2}^q + ...p_{iN}^q)},$$

and the numerator can be expressed as

$$\frac{1}{(N^q - N)}\sum_{i=1}^{S} \sum_{\substack{0 \le r_j < q, j=1,2...,N \\ r_1 + r_2 + ... r_N = q}} \frac{q!}{r_1! r_2! ... r_N!} p_{i1}^{r_1} p_{i2}^{r_2} ... p_{iN}^{r_N}$$

We have a profile of $\{C_{qN}; q = 2, 3, \ldots, N\}$ to describe similarity across N communities. A nearly unbiased estimator of C_{qN} can be constructed by estimating the term $p_{i1}^{r_1} p_{i2}^{r_2} ... p_{iN}^{r_N}$ in the above formula by $X_{i1}^{(r_1)} X_{i2}^{(r_2)} ... X_{iN}^{(r_N)} / (n_1^{(r_1)} n_2^{(r_2)} ... n_N^{(r_N)})$. An approximate variance can be obtained by a similar bootstrap method. See Chao et al. (2008) for details.

Part VII: Genetics

Formulas for Example 7a

All formulas are the same as those in Part VI.