Estimation on the Number of Shared Species via a Jackknife Procedure

Shen, Tsung-Jen
National Chung Hsing University, Department of Applied Mathematics and Institute of Statistics
Taichung (402), Taiwan
City (with Postcode), Country
E-mail: tjshen@nchu.edu.tw

Hwang, Wen-Han
National Chung Hsing University, Department of Applied Mathematics and Institute of Statistics
Taichung (402), Taiwan
E-mail: wenhan@nchu.edu.tw

Chuang, Chia-Jui
National Chung Hsing University, Department of Applied Mathematics and Institute of Statistics
Taichung (402), Taiwan
E-mail: chuangm_philip@yahoo.com.tw

In ecological and biological field surveys, due to cost and time constraints, it is rarely to compile a complete species census in practice. Instead, ecologists and biologists usually attain the same aim with a limited number of samples. Therefore it is important, based on the samples, to gain a reliable inference for the parameter that can characterize ecological communities. This study focuses on the simplest overlap quantity but essential to characterizing two communities — the number of shared species S in two communities.

Assume that there are M_1 species and M_2 species in community I and II respectively. Furthermore, these species can be decomposed as: S shared species in both two communities; $M_1 - S$ species and $M_2 - S$ species unique in community I and II respectively. Let $\mathbf{X} = (X_1, \cdots, X_{M_1})$ and $\mathbf{Y} = (Y_1, \cdots, Y_{M_2})$ be the frequencies by two random samples from the community I and II respectively. Based on the samples, we can only identify those species observed at least once in either sample. If a species is observed in either sample, it is possible to be unique species or shared species. The object of this paper is to estimate the number of shared species S based on such samples.

We denote the relative abundances of species in communities I and II are $(p_1, \ldots, p_{M_1})$ and $(q_1, \ldots, q_{M_2})$, respectively. Let n_1 and n_2 be the sample sizes for two samples (named by sample I and II). Suppose that each individual is observed or detected independently with others, hence the species counts $X_1, \ldots, X_{M_1}$ follow a multinomial distribution with the cell total n_1 and probabilities $p_1, \ldots, p_{M_1}$, and a similar assumption is given to $Y_1, \ldots, Y_{M_2}$ for community II. Let $\mathrm{I}(\cdot)$ be the usual indicator function, when $\mathrm{I}(A) = 1$ if the event A occurs, and 0 otherwise. Denote $D = \sum_{i=1}^{S} \mathrm{I}(X_i \geq 1, Y_i \geq 1)$ the observed number of shared species from the samples. For any two nonnegative integers j and k, we define

$$f_{jk} = \sum_{i=1}^{S} \mathrm{I}(X_i = j, Y_i = k),$$

the number of shared species exactly represented by j and k individuals in two samples. For simplicity, we further let $f_{j+} = \sum_{k\geq 1} f_{jk} = \sum_{i=1}^{S} \mathrm{I}(X_i = j, Y_i \geq 1)$ be the observed number of shared species exactly represented j individuals in sample I, and a similar definition can be applied to f_{+k}. Notice the parameter of interest S can be decomposed into four term below

$$(1) \qquad S = D + f_{0+} + f_{+0} + f_{00},$$

where the last three items are unobservable in samples.

The Jackknife method is invented in Quenouille (1949) and has been widely applied for correcting the statistical bias and the standard error estimation in statistical inference (Shao and Tu, 1995). Regarding ecological issues, Burnham and Overton (1978) applied the procedure to obtain a series of population size estimators for a closed capture-recapture model. Heltshe and Forrester (1983) considered the species richness estimation based on a quadrat sampling data. Traditionally, the first-order jackknife method is carried out through recomputing a desired statistic by successively leaving one observation out at a time from a one-sample set. As our concern is on a two-sample data, we may follow some extended works (Arvesen, 1969; Schechtman and Wang, 2004) for the jackknife procedure regarding two-sample situations to fit the topic of interest from two communities.

Recall that D is the observed number of shared species from the two-sample data, intuitively $\hat{S}_0 = D$ is chosen to be a basic estimator of S for the jackknife procedure. The procedure starts with alternately and sequentially deleting a_ℓ and b_m from the full data and then recounts the observed number of shared species in resulting sub datasets. For instance, $\hat{S}_0^{(-\ell,\cdot)}$ is the observed number of shared species after deleting a a_ℓ from sample I. Trivially, $\hat{S}_0^{(-\ell,\cdot)}$ can be either D or $D-1$ where the later occurs when individual a_ℓ belongs to species i associated with $X_i = 1$ and $Y_i \geq 1$. By performing the usual jackknife method with respect to sample I, it yields an estimator

$$\hat{S}_{0,X} = D + \frac{n_1 - 1}{n_1} f_{1+}.$$

In line with the foregoing arguments and by jackknifing the sample II, we obtain

$$\hat{S}_{0,Y} = D + \frac{n_2 - 1}{n_2} f_{+1}.$$

By taking an weighted average of $\hat{S}_{0,X}$ and $\hat{S}_{0,Y}$ (Arvesen, 1969), our proposed first-order jackknife estimator is

$$(2) \qquad \hat{S}_1 = \frac{n_1 \hat{S}_{0,X} + n_2 \hat{S}_{0,Y}}{n_1 + n_2} = D + \frac{n_1 - 1}{n_1 + n_2} f_{1+} + \frac{n_2 - 1}{n_1 + n_2} f_{+1}.$$

Nevertheless, as shown in Schechtman and Wang (2004), the first-order jackknife estimator does not reduce the bias in terms of asymptotic order. Hence a further correction is necessary. Followed by Schechtman and Wang (2004), we consider to jackknifing $\hat{S}_{0,X}$ with deleting one individual b_m at a time from sample II. As a result, we have derived the second-order estimator $\hat{S}_2$, that is

$$\hat{S}_2 \;=\; D + \frac{n_1 - 1}{n_1} f_{1+} + \frac{n_2 - 1}{n_2} f_{+1} + \frac{(n_1 - 1)(n_2 - 1)}{n_1 n_2} f_{11}.$$

In a brief summary, $\hat{S}_2$ is a consequence by jackknifing $\hat{S}_0$ with alternately deleting one individual from either sample. In order to reducing bias further, we suggest to continue and repeat the procedure stated above. Then we can establish a sequence of estimators $\hat{S}_k, k \geq 0$.

Although the sequence of jackknife estimators $\hat{S}_k$ are likely to gain a smaller bias when k is increased, it inevitably goes with a larger variance. Thus there is a bias-variance trade-off in selecting a jackknife order k. In this study, we consider using a sequential test procedure (Burnham and Overton, 1978) as the decision criterion. For each $k \geq 0$, consider the following testing hypothesis with

$$(3) \qquad H_{0k} : E(\hat{S}_{k+1} - \hat{S}_k) = 0 \text{ vs. } H_{1k} : E(\hat{S}_{k+1} - \hat{S}_k) \neq 0.$$

Suppose that under the null hypothesis H_{0k}, the test statistic

$$(4) \qquad T_k = \frac{\hat{S}_{k+1} - \hat{S}_k}{\sqrt{\widehat{Var}(\hat{S}_{k+1} - \hat{S}_k)}}$$

is asymptotically normal distributed. Given a significance level α, the procedure begins at performing the testing hypothesis of (3) with order $k = 0$ and then continues to the next order until an acceptance occurs. In other words, if the p-value by testing statistic T_k is smaller than α, we consider toward the next hypothesis, and the procedure only stops at the hypothesis associated with the p-value over α first. If the procedure stays at $k = k^*$, our proposed estimator is denoted by $\hat{S}_{JK} = \hat{S}_{k^*}$.

The main result of this study is that the two-sample jackknife procedure in Schechtman and Wang (2004) is extended to estimating shared species richness for two communities. In addition to developing a series of jackknife estimators for the number of shared species, we also suggest a sequential testing criterion for selecting a proper order from these jackknife estimators.

REFERENCES (RÉFERENCES)

- Arvesen, J. N. (1969). Jackknifing U-statistics. *The Annals of Mathematical Statistics*, 40, 2076–2100.
- Burnham, K. P. and Overton, W. S. (1978). Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika*, 65, 3, 625–633.
- Heltshe, J. F. and Forrester, N. E. (1983). Estimating species using the jackknife procedure. *Biometerics*, 39, 1–11.
- Quenouille, M. H. (1949). Approximate tests of correlation in time-series. *Journal of the Royal Statistical Society*, Series B, 11, 68–84.
- Schechtman E. and Wang S. (2004). Jackknifing two-sample statistics. *Journal of Statistical Planning and Inference*, 119, 329–340.
- Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap*, Springer-Verlag, New York

RÉSUMÉ (ABSTRACT)

In this paper, we present a sequence of jackknife estimators for the number of shared species in two communities. With simple and explicit formulae, these estimators are supposed to be computed easily. As the jackknife order increases, the resulting estimator is expected to gain a smaller bias yet is usually accompanied with a larger variance, thus a sequential testing procedure will be proposed to determine the order in applications as well. Several typical methods will be evaluated their performance through a Monte Carlo simulation study. To mimic data sampled from the real world, two forests census datasets established in Malaysia and with 209 shared species between them were considered to be our sampling population.