Ecological Archives E090-073-A1


Appendix A. Statistical derivations.
Appendix A. Statistical derivations.

(1) Construction of a lower confidence limit for the expected minimum number of additional individuals or samples required to detect any fraction (including 100%) of the estimated undetected species

A lower confidence limit for \( m \) would be helpful because of the possible existence of very rare species and large variation of \( m \). We propose using a bootstrap percentile method (Efron and Tibshirani 1993, p. 170) to estimate this confidence limit, as follows. (i) Generate a set of bootstrap sample frequencies \( \{f_0^*, f_1^*, \ldots, f_v^*\} \) (where \( v \) denotes the maximum frequency) from a multinomial distribution with cell total \( n \) and cell probabilities \( \{f_1^*/S_{11}, f_2^*/S_{22}, \ldots, f_v^*/S_{vv}\} \); then compute \( m \) based on the generated bootstrap sample. (ii) Repeat step (i) \( B \) times \( (B \geq 1000 \text{ in most applications}) \) to obtain \( B \) replications \( m_1^*, m_2^*, \ldots, m_B^* \). (iii) The bootstrap percentile method treats the \( B \alpha \)th value in the ordered list of these replications as an approximate \( 1-\alpha \) lower confidence limit. For example, if \( B = 1000 \), then a 95% lower limit is the 50th ordered value of the replications.

(2) Derivational details for replicated incidence case: Equation (14) in the text

Assume that there are \( t \) replicated samples, and in each sample only species presence/absence is recorded. Let \( Q_1 \) and \( Q_2 \) represent the number of species that occur in exactly one sample ("uniques") or in exactly two samples ("duplicates"), respectively.

Define the species detection probability of a species as the chance of encountering at least one individual of the species in any sample. Let the species detection probabilities of the \( S \) species in the community be denoted by \( (p_1, p_2, \ldots, p_s) \) and assume that these probabilities are kept constant across the samples. Let \( X_i \) be the number of samples in which the \( i \)th species was detected. Suppose that \( m \) additional samples are needed to achieve the stopping condition that the expected number of uniques in the enlarged \( t + m \) samples is less than 0.5. Analogous to the derivation of Equation (7) in the text for abundance data, we can obtain that the expected number of uniques in the enlarged \( t + m \) samples:

\[
\sum_{i=1}^{S} (1-p_i)^m I(X_i = 1) + \sum_{i=1}^{S} mp_i(1-p_i)^{m-1} I(X_i = 0) = \frac{E(Q_2)}{t-1}.
\] (A.1)

For incidence data, the Turing formulas need to be modified. First we derive an approximate detection probability for any unique. Consider

\[
E\sum_{i=1}^{S} \frac{p_i}{1-p_i} I(X_i = 1) = \sum_{i=1}^{S} mp_i^2 (1-p_i)^{t-2} = \frac{2E(Q_2)}{t-1}.
\] (A.2)
Assume that all uniques have identical detection probabilities and denote this common probability by $h_1$. Then it follows from (A.2) that

$$h_1 Q_t \approx \frac{2 Q_t}{1 - h_1} t^{-1},$$

which leads to $h_1 = \frac{2 Q_t}{[2 Q_t + (t-1)Q_t]}$. Similarly, consider

$$E \sum_{i=1}^{S} \frac{p_i}{1 - p_i} I(X_i = 0) = \sum_{i=1}^{S} p_i (1 - p_i)^{i-1} = \frac{E(Q_i)}{t}.$$  \hfill (A.3)

Assume that all previously undetected species have identical detection probabilities and denote this common probability by $h_0$. Then it follows from (S3) that

$$h_0 \frac{\hat{Q}_0}{1 - h_0} \approx \frac{Q_i}{t},$$

which yields $h_0 = \frac{Q_i}{[Q_i + t \hat{Q}_0]}$. Thus, Equation (A.1) becomes

$$Q_i (1 - h_1)^m + \hat{Q}_i m h_0 (1 - h_0)^{m-1}. \hfill (A.4)$$

Substituting $h_1 = \frac{2 Q_t}{[2 Q_t + (t-1)Q_t]}$, $h_0 = \frac{Q_i}{[Q_i + t \hat{Q}_0]}$ and the Chao2 estimator $\hat{Q}_o = (1-1/t)Q_t^2 / (2Q_t)$ into (S4), we have (S1) < 0.5 if and only if

$$Q_i (1 + \frac{m}{t}) \left[ 1 - \frac{2 Q_t}{(t-1)Q_t + 2Q_t} \right]^m < 0.5 \hfill (A.5)$$

Using the approximation $[1 - \frac{2 Q_t}{(t-1)Q_t + 2Q_t}]^m \approx \exp\left[ - \frac{m}{t} \frac{2Q_t}{(t-1)Q_t + 2Q_t} \right]$, we then obtain Equation (14) in the text.

(3) Derivational details for replicated incidence case: Equation (15) in the text

If $g$ is the fraction of $S_{est}$ that is desired ($0 < g < 1$), then the objective is to find the additional $m_g$ samples such that the number of species reaches the target value $g S_{est}$, i.e., the number of previously undetected species discovered in the additional sample is $g S_{est} - S_{obs}$. The expected number of previously undetected species in additional $m_g$ samples can be obtained as

$$\sum_{i=1}^{S} [1 - (1 - p_i)^m] I(X_i = 0).$$ \hfill (A.5)

Assume that all previously undetected species have identical detection probabilities and denote this common probability by $h_0$. We have derived that $h_0 = \frac{Q_i}{[Q_i + t \hat{Q}_0]}$. Thus
\[ (S5) \approx \hat{Q}_0 [1 - (1 - h_0)^{m^*}] \approx g \, S_{est} - S_{obs}. \]

Substituting \( h_0 = Q_t / (Q_t + t\hat{Q}_0) \) and \( \hat{Q}_0 = (1 - 1/t)Q_t^3 / (2Q_2) \) into the above and solving for \( m_g \), we then obtain Equation (15) in the text.

**Literature cited**