

PENNSSTATE



---

## Center for Statistical Ecology and Environmental Statistics

---

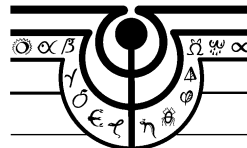
### ESTIMATION OF SPECIES RICHNESS BASED ON SPECIES RANGE

By G. P. Patil and C. Taillie

Center for Statistical Ecology and Environmental Statistics  
Department of Statistics  
The Pennsylvania State University  
University Park, PA 16802

[Invited Paper to appear in *Community Ecology*, 2002]

Technical Report Number 951201  
TECHNICAL REPORTS AND REPRINTS SERIES  
December 2001



---

Department of Statistics  
The Pennsylvania State University  
University Park, PA 16802

G. P. Patil  
Distinguished Professor and Director  
Tel: (814)865-9442 Fax: (814)865-1278  
Email: [gpp@stat.psu.edu](mailto:gpp@stat.psu.edu)  
<http://www.stat.psu.edu/~gpp>

# Estimation of Species Richness Based on Species Range

G. P. Patil<sup>1</sup> and C. Taillie<sup>1</sup>

<sup>1</sup>Center for Statistical Ecology and Environmental Statistics, Department of Statistics, Pennsylvania State University, University Park, PA 16802

**Keywords:** Expansion estimator, Horvitz-Thompson estimator, Hypergeometric probability, Ill-posed problems, Inclusion probability, Quadrat sampling, Regularization.

**Abstract:** A geographical region, containing an unknown number  $S$  of species, is partitioned into  $N$  quadrats. The *range* of a species is defined to be the number of quadrats in which the species is present. A random sample of  $n$  quadrats is drawn without replacement, and the species list is determined for each of the selected quadrats. Two estimators of  $S$  are proposed. The inclusion probabilities in the Horvitz-Thompson estimator involve the unknown species ranges but these ranges can be estimated to yield an “estimated” Horvitz-Thompson estimator. This estimator is biased because of the use of estimated inclusion probabilities. For the other estimator, it is shown that the expected number of species in the sample having a specified sample range  $r$  is a linear combination over  $R$  of the number  $S_R$  of species in the population with population range  $R$ . Letting  $r$  vary yields a system of linear equations that can be solved to obtain estimates for the  $S_R$  and for  $S$ . These estimators for  $S_R$  and  $S$  are shown to be unbiased when the sample size  $n$  is sufficiently large.

## 1. Introduction

A geographical region, containing an unknown number  $S$  of species, is partitioned into  $N$  quadrats. Typically, the quadrats would be of equal area, but this is not required. A random sample of  $n$  quadrats is chosen without replacement, and the species list is determined (without error) for each of the selected quadrats. It is desired to estimate  $S$  on the basis of these  $n$  species lists.

Here, we suggest two possible estimators but we do not enter into a detailed examination of their behavior. The first estimator is based upon a classification of species according to their range, where the range of a species is here taken to be the number of quadrats in which the species is present. To see why “range” is relevant, consider the situation in which each species has its range equal to unity, i.e., each species occurs in exactly one quadrat although a given quadrat may contain many species. In this case, the species richness variable is additive across quadrats and the simple expansion estimator,

$$\hat{S} = \frac{N}{n} s, \quad (1)$$

where  $s$  number of species in the sample, is unbiased. Simple expansion does not work when a species has range greater than unity since the presence of the species in additional quadrats does not increase total species richness but does change the inclusion probability for the species.

For our first estimator, the method of moments is applied to estimate the number of species in the population that have a given range  $R$ . These estimates are then summed across the possible values of  $R$  to obtain the estimate of species richness. The resulting estimator is shown to have two properties: (i) it is unbiased provided

$n$  is greater than or equal to the range of every species in the population, and (ii) it reduces to the expansion estimator (1) when every species has its range equal to unity.

Our second estimator is of the Horvitz-Thompson type with species as sampling unit. The inclusion probability for a species depends upon its unknown range so we estimate the inclusion probability by estimating the range.

Throughout we use capital letters for population quantities and lower case letters for the corresponding sample values.

## 2. Moment Estimator

Let  $S_R, R = 1, \dots, N$ , be the set of species in the population whose range is  $R$  and write  $S_R$  for the cardinality of  $S_R$ . Since  $S_1, S_2, \dots, S_N$  partition the set of species, we have

$$S = S_1 + S_2 + \dots + S_N.$$

The estimator in this section works best when all (or at least most) of the  $S_R$  vanish for  $R$  sufficiently large.

Define the sample range  $r$  of a species as the number of sampled quadrats in which the species occurs, and put  $s_r, r = 1, 2, \dots, n$  equal to the number of species whose sample range is  $r$ . Just as for the population quantities we have

$$s = s_1 + s_2 + \dots + s_n.$$

The proposed estimator is based upon the following result:

$$E[s_r] = \sum_{R=r}^N \frac{\binom{N-R}{n-r} \binom{R}{r}}{\binom{N}{n}} S_R, \quad r = 1, 2, \dots, n. \quad (2)$$

The coefficient involving the binomial coefficients is the (hypergeometric) probability that a given species with population range  $R$  has its sample range equal to  $r$ . These probabilities are summed over all species in population to obtain equation (2).

For estimation we replace  $E[s_r]$  by  $s_r$  in equation (2) and write  $H_{rR}$  for the hypergeometric probability, giving the system of linear equations

$$s_r = \sum_{R=1}^N H_{rR} S_R, \quad r = 1, 2, \dots, n. \quad (3)$$

to be solved for the  $S_R$ . This system of equations has more unknowns than equations, so that a unique solution does not exist. One possible solution is obtained by setting  $S_{n+1} = S_{n+2} = \dots = S_N = 0$  and to arrive at a nonsingular system

$$s_r = \sum_{R=1}^n H_{rR} S_R, \quad r = 1, \dots, n. \quad (4)$$

of  $n$  equations in  $n$  unknowns. This system is easily solved by backward substitution since the matrix of coefficients is upper triangular. Writing the solution of (4) as  $\hat{S}_1, \hat{S}_2, \dots, \hat{S}_n$ , the moment estimator of species richness based on species range becomes

$$\hat{S}_{trunc} = \hat{S}_1 + \hat{S}_2 + \dots + \hat{S}_n. \quad (5)$$

We refer to the estimator (5) as the *truncated* moment estimator because the parameters  $S_R$  have been set to zero for  $R > n$ . Due to linearity, the estimator (5) is unbiased provided it is in fact true that  $S_R = 0$  for  $R = n + 1, \dots, N$ . More generally, the bias is given as

$$\text{bias} = [U_n H^{-1} B - U_{N-n}] S^{(2)} \quad (6)$$

where  $U_t$  and  $t$ -dimensional row vector with unity in each component,  $H$  is an  $n$  by  $n$  matrix consisting of the first  $n$  columns of the coefficient matrix in equation (3),  $B$  is  $n$  by  $n - N$  and consists of the last  $N - n$  columns in the coefficient matrix of equation (3), and  $S^{(2)}$  is a column vector containing  $S_{n+1}, S_{n+2}, \dots, S_N$ .

As a special case, suppose  $S_R$  vanishes for  $R \geq 2$ , so that each species has unit range. Then, backward substitution shows that  $S_R$  also vanishes for  $R \geq 2$  and

$$\hat{S}_1 = \frac{N}{n} s_1.$$

The truncated moment estimator (5) then reduces to the sample expansion estimator (1).

It is possible to obtain a very complicated expression for the variance of  $\hat{S}_{trunc}$ . But the formula involves the second order joint occurrence probabilities and it would appear to be difficult to estimate the variance using this formula. Some type of resampling may be more suitable for variance estimation.

A major open question is providing data-based guidance as to whether and when  $n$  is large enough. It may happen that the tail of the sequence  $S_1, S_2, \dots, S_N$  vanishes except for a few isolated nonzero values corresponding to a few spatially abundant species. If sampling is sufficiently intense, all these species should occur in the sample and produce some isolated nonzero values in the tail of the sequence  $s_1, s_2, \dots, s_n$ . Richness can be estimated by removing these species, and applying the truncated moment estimator to the remaining species.

The truncated moment estimator would appear to be inappropriate when numerous species have large ranges. In this case, estimation is an ill-posed problem since the estimation equations (3) involve more unknown parameters than there are equations. One way of handling this situation is through the method of regularization, which estimates the parameters  $S_1, S_2, \dots, S_N$  by minimizing the expression

$$\sum_{r=1}^n \left( s_r - E[s_r] \right)^2 + \lambda Q(S_1, S_2, \dots, S_N), \quad (7)$$

where  $\lambda$  is a smoothing parameter and where  $Q$  is a quadratic function that imposes a penalty for lack of smoothness of the sequence  $S_1, S_2, \dots, S_N$ . Vapnik (1995, p. 9) gives a succinct one-page description of the regularization method. See Press *et al* (1992, chap. 18) for discussion of possible choices of  $Q$ . We refer to the resulting estimator of species richness as the *regularized* moment estimator. Note that the regularized estimate is not generally unbiased and its value depends upon the choice of  $Q$  and  $\lambda$ . There is reason to expect that the latter dependence is not overly strong since the sum  $S_1 + S_2 + \dots + S_N$  should not change much when the sequence  $S_1, S_2, \dots, S_N$  is smoothed. This question has not been examined, however.

### 3. Estimated Horvitz-Thompson Estimator

With unequal probability sampling, the Horvitz-Thompson estimator of population size is the sum of the reciprocals of the inclusion probabilities where the sum is over all units entering the sample. This estimator is unbiased.

In our situation, the sampling design is equal probability on quadrats; but it induces an unequal probability sampling on the set of species. Species richness can be estimated as described in the previous paragraph as long as we can determine the inclusion probabilities for the different species. But, if a species has population range  $R$ , then its inclusion probability is seen to be

$$\pi = 1 - \frac{\binom{N-R}{n}}{\binom{N}{n}}. \quad (8)$$

The difficulty is that we do not know the value of  $R$ . But it can be estimated and the expansion estimator of  $R$  is

$$\hat{R} = \frac{N}{n} r \quad (9)$$

where  $r$  is the sample range of the species in question. One may wish to round  $\hat{R}$  to an integer, but we do not introduce any special notation for this. The corresponding estimate for the inclusion probability is

$$\hat{\pi} = 1 - \frac{\binom{N-\hat{R}}{n}}{\binom{N}{n}} \quad (10)$$

and the estimated Horvitz-Thompson estimator becomes

$$\begin{aligned}
\hat{S}_{HT} &= \sum_{\text{sample}} \frac{1}{\hat{\pi}} \\
&= \sum_{r=1}^n \frac{s_r}{\binom{N-\hat{R}}{n}} \\
&\quad 1 - \frac{\binom{N}{n}}{\binom{N-\hat{R}}{n}}
\end{aligned} \tag{11}$$

The properties of this estimator have not been investigated.

## References

Press, W. H, Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). *Numerical Recipes In C*, second edition. Cambridge University Press.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer, New York.