

Inferring Species Richness from Samples

Paul Murtaugh

Department of Statistics
Oregon State University

This publication presentation was developed under STAR Research Assistance Agreement No. CR82-9096-01 awarded by the U.S. Environmental Protection Agency to Oregon State University. It has not been formally reviewed by EPA. The views expressed in this document are solely those of the author, and EPA does not endorse any products or commercial services mentioned in this publication.

THE SITUATION

Suppose we have a community of S species, from which we take a sample of n individuals.

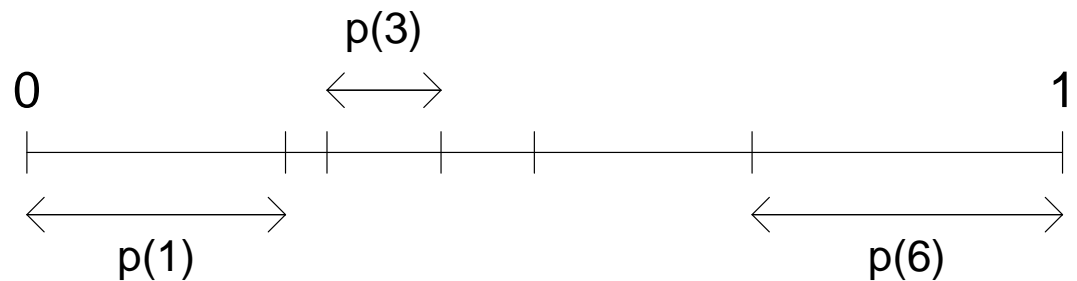
If $S_n \leq S$ is the number of species observed in the sample, what is our best estimate of S ?

The answer depends on

- the portion of the community included in the sample; and
- the relative abundance of the S species in the community.

THE BROKEN-STICK MODEL

Take a random sample of size $S - 1$ from $U[0,1]$:



Relative abundances of S species are given by the segment lengths.

THE MULTINOMIAL MODEL

Given true relative abundances $(\pi_1, \pi_2, \dots, \pi_S)$, the counts of sampled individuals in the different species comprise a multinomial random vector:

$$(n_1, n_2, \dots, n_S) \sim \text{Multinomial}(n; \pi_1, \dots, \pi_S),$$

with $\sum n_i = n$.

If $S_n < S$, then some of the n_i 's will be zero.

STRATEGY

1. Observe $(n_1, n_2, \dots, n_{S_n})$, with $n_1 < n_2 < \dots < n_{S_n}$.
2. For $S = S_n, S_n + 1, S_n + 2, \dots$,
 - (a) Add $S - S_n$ zeroes to the vector of observed counts: $(0, \dots, 0, n_{S-S_n+1}, \dots, n_S)$.
 - (b) Do a χ^2 goodness-of-fit test comparing the observed counts to those expected under the broken-stick model $(n\pi_1, n\pi_2, \dots, n\pi_S)$.
3. Choose the value of S giving the largest GOF P -value, and construct a CI as the set of all S 's for which $P > 0.05$.

PRELIMINARY RESULTS & ISSUES

The method works well for data simulated with the broken-stick model, but is not at all robust to other possible models of relative abundance.

Other ideas:

- Use information on the *change* in species richness with increasing sample size (*rarefaction*).
- Develop Bayesian approaches that start with priors on S and the vector of relative abundances.