

Statistical Estimation of Uncultivated Microbial Diversity

J. Bunge (✉)

1	Introduction	2
2	Abundance Data	3
2.1	Parametric Abundance Models	4
2.2	Nonparametric Abundance Models	7
2.3	Coverage-Based Estimation	8
2.4	Discussion	9
3	Incidence Data	10
3.1	Parametric Incidence Models	12
3.2	Nonparametric Incidence Models	13
3.3	Nonparametric Coverage-Based Incidence Methods	14
3.4	Models with Covariates	14
4	Remarks	14
	APPENDIX: Software	16
	References	17

Abstract The full microbial richness of a community, or even of an environmental sample, usually cannot be observed completely, but only estimated statistically. This estimation is typically based on observed count data, that is, the counts of the representatives of each species (or other taxonomic units) appearing in the sample or samples. “Abundance” data consists of counts of the numbers of individuals from various species in a single sample, while “incidence” (or multiple recapture) data consists of lists of species appearing in several or many samples. In this chapter we consider statistical estimation of the total richness, i.e., the total number of species, observed + unobserved, based on abundance or on incidence data. We discuss parametric and nonparametric methods, their underlying assumptions, and their advantages and disadvantages; computational implementations and software; and larger scientific issues such as the scope of applicability of the results of a given analysis. Some real-world examples from microbial studies are presented. Our discussion is intended to serve as an overview and an introduction to the literature and available software.

J. Bunge

Department of Statistical Science, Cornell University, Ithaca, NY 14853, USA
e-mail: jab18@cornell.edu

1 Introduction

Recent research has shown that microbial communities are astonishingly diverse; in fact many studies only capture a small fraction of the diversity of a given community, despite intensive sampling efforts (Huber et al. 2007). In such cases we must estimate the total diversity – observed plus unobserved – by statistical extrapolation from the available data. This is a nontrivial and indeed not entirely solved problem in statistics; it is a topic of considerable interest and activity among theoretical (mathematical) statisticians, and its literature continues to evolve at a rapid rate (Bunge and Barger 2008). Some of these statistical developments have entered the mainstream of microbial diversity research, but some have not. In this chapter we give an overview of the area from an applied, data-analytic perspective, with the goal of providing the practitioner with a conceptual framework for the diversity estimation problem; the types of data typically encountered; and the relevant statistical procedures that are applicable to such datasets.

First we require a definition of “diversity.” This in turn requires that the community or population in question be classified in a clear and unambiguous manner, i.e., that it be subdivided into mutually exclusive subsets which, together, comprise the entire population. For statistical purposes any well-defined classification system will do, but in a biological population it is natural to classify individual organisms according to the Linnaean hierarchy, in particular by “species.” However, there is currently no consensus regarding the concept of “species” for microorganisms, and instead microbiologists often group environmental microorganisms into operational taxonomic units (OTUs) based on their rRNA gene sequence similarity (Stackebrandt and Goebel 1994). A species is then provisionally defined to be a group (OTU) of cells sharing a certain percentage identity of their 16s rRNA gene sequences. Values from 97 to 99% are typically used.

Given a classification system, several indices of diversity have been defined (Magurran 2004), but the simplest is the number of OTUs, or “species richness” in a given community. This quantity has a clear physical meaning and in principle could be determined exactly, given unlimited sampling effort. However, species richness, while relatively straightforward to define, is difficult to ascertain in practice, because biological communities often comprise a few large and many small species, and it is precisely the small species that elude sampling efforts. That is, the unobserved part of the community may be subdivided into many small groups unknown to us, yet we are required to estimate the number of these unobserved species. This is why the statistical problem of estimating species richness does not at present have an optimal, universal solution. Indeed, some authors have argued (mathematically) that no such global solution is possible, and that under the most general, nonparametric formulation of the problem one can at best provide a lower bound for the species richness of a given population (Mao and Lindsay 2007). On the other hand, if one is willing to impose certain structural constraints, richness estimation becomes possible, although subject to the validity of the assumed structure. For this and other reasons it is advisable to use and compare several existing methods, which make different assumptions about the (unknown) structure of the population.

The goal of this chapter is not to comprehensively review the current literature or practice (statistical or biological), but to describe the scope and applicability of the major statistical methods from a synoptic, and somewhat idealized, perspective. (In particular, the references given here are intended as entry points to the literature not definitive historical summaries.) This is because the status of current theory and practice are, to a certain degree, fragmented and incomplete. The various methods have not yet been unified in a single mathematical framework, and in particular there is no comprehensive expository textbook, at the theoretical or applied level. More importantly from the practitioner's point of view, there is no unified and comprehensive software program for species richness estimation. Some methods have been implemented in software that can be readily used by the applied practitioner, others in software that requires a statistical computing specialist, while for others no software exists at all. In this chapter we seek to give an overview of the state of the art. We take a broad perspective, attempting to look beyond the present limitations of the literature or software resources (which at any rate are being continually improved), while referring the reader to current and relevant existing resources where possible. We focus on those methods for which the mathematical foundations have been studied in depth.

Generally speaking, two types of data are encountered in species richness estimation: first, abundance or frequency count data, usually from a single sample; and second, incidence or occurrence data, usually from multiple samples (from the same community). In the next two sections we discuss statistical methods for each of these data types, and connections between them. In the final section, we discuss certain scientific issues (not purely statistical) and potential future directions.

2 Abundance Data

In this scenario we collect a sample of organisms, sort them into species, and count how many of each kind we have in the sample. Such a description hides the complications of the data-collection process, which may have several stages, each with its own biases (as in the case of clone library construction), and it hides the somewhat arbitrary decisions underlying the operative definition of species or OTU. However, the procedure is at least conceptually clear, and we will relegate its uncertainties to the background for now, in order to focus on the statistical methods.

Given such a sample, then, how can we interpret it statistically? Since the total species list is unknown (otherwise there would be no estimation problem), there is no obvious ordering of the species observed in the sample. We therefore organize the data by simply counting the number of species observed once (the “singletons”), twice, three times, and so on. For example, in the dataset (1,25), (2,7), (3,7), (4,4), (5,1), (6,2), (8,1), (11,1), (13,1), (14,1), (16,1), (27,1), (31,1), and (37,1); there were 25 species observed once (each), seven observed twice, seven observed three times, ..., and 1 observed 37 times (example data from Behnke et al. (2008); OTUs defined at 98% sequence similarity level). Thus there were

Uncultivated Microorganisms

Epstein, S.S. (Ed.)

2009, X, 208 p. 33 illus., 10 illus. in color., Hardcover

ISBN: 978-3-540-85464-7