Microbiology Monographs
Volume 10

Series Editor: Alexander Steinbüchel
Münster, Germany

Microbiology Monographs

Volumes published in the series

Inclusions in Prokaryotes
Volume Editor: Jessup M. Shively
Vol. 1, 2006

Complex Intracellular Structures in Prokaryotes
Volume Editor: Jessup M. Shively
Vol. 2, 2006

Magnetoreception and Magnetosomes in Bacteria
Volume Editor: Dirk Schüler
Vol. 3, 2007

Predatory Prokaryotes – Biology, Ecology and Evolution
Volume Editor: Edouard Jurkevitch
Vol. 4, 2007

Amino Acid Biosynthesis – Pathways, Regulation and Metabolic Engineering
Volume Editor: Volker F. Wendisch
Vol. 5, 2007

Molecular Microbiology of Heavy Metals
Volume Editors: Dietrich H. Nies and Simon Silver
Vol. 6, 2007

Microbial Linear Plasmids
Volume Editors: Friedhelm Meinhardt and Roland Klassen
Vol. 7, 2007

Prokaryotic Symbionts in Plants
Volume Editor: Katharina Pawlowski
Vol. 8, 2009

Hydrogenosomes and Mitosomes: Mitochondria of Anaerobic Eukaryotes
Volume Editor: Jan Tachezy
Vol. 9, 2008

Uncultivated Microorganisms
Volume Editor: Slava S. Epstein
Vol. 10, 2009

Microbial Megaplasmids
Volume Editor: Edward Schwartz
Vol. 11 2009

Endosymbionts in Paramecium
Volume Editor: Masahiro Fujishima
Vol. 12, 2009

Alginates: Biology and Applications
Volume Editor: Bernd H. A. Rehm
Volume 13, 2009

Slava S. Epstein
Editor

Uncultivated Microorganisms

Springer

Editor
Dr. Slava S. Epstein
Department of Biology
134 Mugar Hall
Northeastern University
360 Huntington Ave
Boston MA 02115
USA
e-mail: slava.epstein@gmail.com

Series Editor
Professor Dr. Alexander Steinbüchel
Institut für Molekulare Mikrobiologie und Biotechnologie
Westfälische Wilhelms-Universität
Corrensstraße 3
48149 Münster
Germany
e-mail: steinbu@uni-muenster.de

ISSN 1862-5576 e-ISSN 1862-5584
ISBN 978-3-540-85464-7 e-ISBN 978-3-540-85465-4
DOI: 10.1007/978-3-540-85465-4
Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: 2009929709

Cover design: SPi Publisher Services

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

In 1898, an Austrian microbiologist Heinrich Winterberg made a curious observation: the number of microbial cells in his samples did not match the number of colonies formed on nutrient media (Winterberg 1898). About a decade later, J. Amann quantified this mismatch, which turned out to be surprisingly large, with non-growing cells outnumbering the cultivable ones almost 150 times (Amann 1911). These papers signify some of the earliest steps towards the discovery of an important phenomenon known today as the Great Plate Count Anomaly (Staley and Konopka 1985). Note how early in the history of microbiology these steps were taken. Detecting the Anomaly almost certainly required the Plate. If so, then the period from 1881 to 1887, the years when Robert Koch and Petri introduced their key inventions (Koch 1881; Petri 1887), sets the earliest boundary for the discovery, which is remarkably close to the 1898 observations by H. Winterberg. Celebrating its 111th anniversary, the Great Plate Count Anomaly today is arguably the oldest unresolved microbiological phenomenon.

In the years to follow, the Anomaly was repeatedly confirmed by all microbiologists who cared to compare the cell count in the inoculum to the colony count in the Petri dish (*cf.,* Cholodny 1929; Butkevich 1932; Butkevich and Butkevich 1936). By mid-century, the remarkable difference between the two counts became a universally recognized phenomenon, acknowledged by several classics of the time (Waksman and Hotchkiss 1937; ZoBell 1946; Jannasch and Jones 1959).

Surely the "missing" microbial diversity was as large then as it is now. However, reading the earlier papers leaves an impression that throughout most of the 20th century the "missing" aspect was not viewed as a particularly important problem or as an exciting opportunity. A casual mention was typical of many publications. "Missing" cells were not necessarily considered missing species let alone signs of novel classes of microbes. Besides, the unexplored microbial biodiversity was a purely academic issue; the hunt for novel species as a resource for biotechnology had not yet begun. It is also important that the reasons for the Anomaly appeared rather simple at the time. Counting errors, dead cells, and later damaged cells were continuously considered significant components of the disparity. Also, it had been obvious at least since Koch's time that no single nutrient medium could possibly satisfy all microorganisms (Koch 1881), and so the finger was always pointing to media deficiencies. Indeed, imperfections in media design was such a simple and intuitive explanation for the refusal of the microbial majority to grow *in vitro* that many microbiologists began viewing it as sufficient. The triviality of the explanation generated a perception of the Anomaly as a purely technical issue that could be resolved by bettering the media compositions and incubation conditions.

This view began to change towards the end of the 20th century. Cultivation efforts during the preceding decades did produce success stories; yet even as the manuals for media recipes grew into thick volumes, the overwhelming majority of microorganisms still eschewed the Petri dish. The progress in recovering missing species was rather incremental and did not change the overall picture. And, it was going to get worse.

The rRNA approach (Olsen et al. 1986) was a truly spectacular development: it provided insight into the microbial world missed by traditional cultivation. Novel microbial divisions were discovered by the dozen (Giovannoni et al. 1990; Ward et al. 1990; DeLong 1992; Fuhrman et al. 1992; Liesack and Stackebrandt 1992; Barns et al. 1994; Hugenholtz et al. 1998; Ravenschlag et al. 1999; Dojka et al. 2000). From the molecular surveys of the 1990s emerged an image of the biosphere with millions of novel microbial species waiting to be discovered (Tiedje 1994; Allsopp et al. 1995). What microbiologists had been able to cultivate and catalogue throughout the entire history of microbiological exploration (Staley et al. 1989) appeared to be an insignificant portion of the total. Successes in cultivation notwithstanding, the gap between microbial richness in nature and that of culture collections just would not close. Even today, most of the known microbial divisions have no single cultivable representative (Rappe and Giovannoni 2003; Schloss and Handelsman 2004). This gap was called "extraordinary" in 1932 just as it was called in 2000 (Butkevich 1932; Colwell 2000), as if the countless cultivation studies during these seventy years never existed. But, the realities of our age are different from the 1930s, and the Great Plate Count Anomaly is no longer "just" an academic observation. The need to close the gap is an urgent practical issue, as biotech and pharmaceutical industries appear to have exhausted what the limited number of cultivable species have to offer (Osburne et al. 2000). Today, the resolution of the phenomenon of microbial uncultivability is recognized as a top research priority for microbial biology (Young 1997; Hurst 2005). The principal challenges are to understand why uncultivated microorganisms are uncultivated, and to describe, access, and utilize their seemingly infinite diversity.

Microbiologists answered the call using two different strategies. One represents a group of clever approaches that bypass cultivation altogether. These go straight to the genes of the "missing" species to mine them for the information and products they encode, or employ isotopes and miniature electrodes to measure the activities of these species *in situ*. It is truly exciting to see how, today, cultivation-independent studies can be done at a single cell level. The other is a head-on strategy, and consists of a multitude of innovations in cultivation, principally aiming at mimicking natural conditions. The two strategies have their specific advantages and disadvantages, but few microbiologists think it is a battle of two competing products. Instead, the likely solution to the Anomaly is in a symbiosis between the two. What form and shape this symbiosis will take, it is too early to say, but the good news is that both the authors and the readers of this book will likely witness, and witness soon, the process and the conclusion of this evolution.

Furthering this unification is the main goal of this volume. The contributions center around three themes. The first theme groups together several chapters that focus on what can be learned about the microbial world without cultivating it. John Bunge opens the volume by describing how to statistically estimate the size of microbial diversity using gene sequence data. Chapters by Mitchell Sogin and Terry Gentry et al. provide an account of the state of the art in recovering the sequence data from environmental samples. Antje Boetius et al. offer a perspective on studying microbes in nature by measuring their biogeochemical activities. Mircea Podar et al. explore how much information modern genomics tools can recover from single cells of uncultivated species. The second theme is the nature of uncultivated microorganisms, why so many species remain uncultivated, and how to domesticate them in the lab. Thomas Schmidt and Allan Konopka dissect the nature of slow growing species. Rita Colwell describes cells that are viable but nonculturable. Slava Epstein attempts to build a general model of the Great Plate Count Anomaly. The third theme builds connections between the Anomaly and practice. Vivian Miao and Julian Davies explore how metagenomics approaches could help to provide access to bioactive compounds produced by uncultivated species. Kim Lewis advocates a connection between the phenomemon of microbial uncultivability and antimicrobial tolerance of biofilms and persister cells. Ken Nealson's chapter concludes the volume by discussing the application of uncultivated cells to the search of life outside our planet, so as to outline the wide spectrum of subjects connected to the Great Plate Count Anomaly.

There is something else that this volume intends to convey. Working in the field of uncultivated microorganisms today involves both luck and privilege. Not everyone has the fortune to study a phenomenon that has endured for a century, is of unquestionable importance, and yet remains unresolved. It is fascinating to think that "our" phenomenon predates the model of the atom, the theory of the Big Bang, cracking the genetic code… It is humbling to think of the great minds who have contributed over the past century to its resolution. And it is sensational to think how enormously beneficial this resolution may be by providing unimpeded access to the missing microbial diversity, and the treasures therein. As to the luck … the luck is in the timing, for the right of entry into the world of uncultivated microbes seems to be just round the corner.

Boston MA
January 2009

Slava S. Epstein

References

Allsopp D, Colwell RR, Hawksworth DL (1995) Microbial diversity and ecosystem function. CAB International, Wallingford, UK

Amann J (1911) Die direkte Zählung der Wasserbakterien mittels des Ultramikroskops. Centralbl f Bakteriol 29:381–384

Barns SM, Fundyga RE, Jeffries MW, Pace NR (1994) Remarkable archaeal diversity detected in a Yellowstone National Park hot spring environment. Proc Natl Acad Sci USA 91:1609–1613

Butkevich NV, Butkevich VS (1936) Multiplication of sea bacteria depending on the composition of the medium and on temperature. Microbiology 5:322–343

Butkevich VS (1932) Zür Methodik der bakterioloschen Meeresuntersuchungen und einige Angaben über die Verteilung der Bakterien im Wasser und in den Büden des Barents Meeres. Trans Oceanogr Inst Moscow 2:5–39 (in Russian with German summary)

Cholodny N (1929) Zur Methodik der quantitativen Erforschung des bakteriellen Planktons. Zentralbl Bakteriol Parasitenkd Infektionskr Hyg A 77:179–193

Colwell RR, Grimes DJ (2000) Nonculturable microorganisms in the environment. ASM Press, Washington DC

DeLong EF (1992) Archaea in coastal marine environments. Proc Natl Acad Sci USA 89:5685–5689

Dojka MA, Harris JK, Pace NR (2000) Expanding the known diversity and environmental distribution of an uncultured phylogenetic division of bacteria. Appl Environ Microbiol 66:1617–1621

Fuhrman JA, McCallum K, Davis AA (1992) Novel major archaebacterial group from marine plankton. Nature 356:148–149

Giovannoni SJ, Britschgi TB, Moyer CL, Field KG (1990) Genetic diversity in Sargasso Sea bacterioplankton. Nature 345:60–63

Hugenholtz P, Goebel BM, Pace NR (1998) Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. J Bacteriol 180:4765–4774

Hurst CJ (2005) Divining the future of microbiology. ASM News 71:262–263

Jannasch HW, Jones GE (1959) Bacterial populations in seawater as determined by different methods of enumeration. Limnol Oceanogr 4:128–139

Koch R (1881) Zur Untersuchung von pathogenen Organismen. Mitth a d Kaiserl 1:1–48

Liesack W, Stackebrandt E (1992) Occurrence of novel groups of the domain Bacteria as revealed by analysis of genetic material isolated from an Australian terrestrial environment. J Bacteriol 174:5072–5078

Olsen GJ, Lane DJ, Giovannoni SJ, Pace NR, Stahl DA (1986) Microbial ecology and evolution: a ribosomal RNA approach. Annu Rev Microbiol 40:337–365

Osburne MS, Grossman TH, August PR, MacNeil IA (2000) Tapping into microbial diversity for natural products drug discovery. ASM News 66:411–417

Petri RJ (1887) Eine kleine Modification des Koch'schen Plattenverfahrens. Centralbl f Bakteriol 1:279–280

Rappe MS, Giovannoni SJ (2003) The uncultured microbial majority. Annu Rev Microbiol 57:369–394

Ravenschlag K, Sahm K, Pernthaler J, Amann R (1999) High bacterial diversity in permanently cold marine sediments. Appl Environ Microbiol 65:3982–3989

Schloss PD, Handelsman J (2004) Status of the microbial census. Microbiol Mol Biol Rev 68:686–691

Staley JT, Bryant MP, Pfennig N, Holt JG (1989) Bergey‘s manual of sytematic bacteriology

Staley JT, Konopka A (1985) Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. Annu Rev Microbiol 39:321–346

Tiedje JM (1994) Microbial diversity: of value to whom? ASM News 60:524–525

Waksman SA, Hotchkiss M (1937) Viability of bacteria in sea water. J Bacteriol 33:389–400

Ward DM, Weller R, Bateson MM (1990) 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. Nature 345:63–65

Winterberg H (1898) Zur Methodik der Bakterienzahlung. Zeitschr f Hyg 29:75–93

Young P (1997) Major microbial diversity initiative recommended. ASM News 63:417421

ZoBell CE (1946) Marine microbiology: a monograph on hydrobacteriology. Chronica Botanica Co, Waltham, MA, USA

Contents

Statistical Estimation of Uncultivated Microbial Diversity 1
J. Bunge

Characterizing Microbial Population Structures through Massively Parallel Sequencing 19
Mitchell L. Sogin

Detection and Characterization of Uncultivated Microorganisms Using Microarrays 35
Terry J. Gentry, Zhili He, and Jizhong Zhou

The Seabed as Natural Laboratory: Lessons From Uncultivated Methanotrophs 59
Antje Boetius, Thomas Holler, Katrin Knittel, Janine Felden, and Frank Wenzhöfer

Single Cell Whole Genome Amplification of Uncultivated Organisms 83
Mircea Podar, Martin Keller, and Philip Hugenholtz

Physiological and Ecological Adaptations of Slow-Growing, Heterotrophic Microbes and Consequences for Cultivation 101
Thomas M. Schmidt and Allan E. Konopka

Viable but Not Cultivable Bacteria 121
Rita R. Colwell

General Model of Microbial Uncultivability 131
Slava S. Epstein

Metagenomics and Antibiotic Discovery from Uncultivated Bacteria 161
Vivian Miao and Julian Davies

Persisters, Biofilms, and the Problem of Cultivability 181
Kim Lewis

Taking the Concept to the Limit: Uncultivable Bacteria and Astrobiology 195
Kenneth Nealson

Index 205

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

Microbiol Monogr, doi:10.1007/7171_2008_3

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 unbeknown 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

25 + 7 + 7 + 4+…+1 = 54 species observed altogether in this sample, and there were 1 × 25 + 2 × 7 + 3 × 7 + … + 37 × 1 = 250 individuals.

If we denote such a dataset in general by $\{(i,f_i),i = 1,2,\ldots\}$, then i is the *frequency* (of sample occurrence) and f_i is the *frequency of the frequency i* (in the sample), an unwieldy phrase which we may replace by *frequency count*, and we call this *frequency count data*. It has a simple and intuitive graphical display with frequency on the horizontal axis and count on the vertical axis, as shown for our example data in Fig. 1.

The left-hand side of the graph represents the less-abundant or rare species (at least in terms of their representation in the observed sample), and the right-hand side represents the abundant or frequent species (in the sample). Note that the structure shown in Fig. 1 is typical of microbial diversity studies: almost half of the observed species (25/54) were represented by singletons in the sample, but at the same time there were three highly abundant species (with frequencies 27, 31, and 37), each accounting for more than 10% of the sampled individuals (27/250, 31/250, 37/250). This structure reflects the statement noted in the Introduction, that the unobserved portion of the population may be subdivided to an almost arbitrary degree, rendering species-richness estimation difficult, or at least prone to statistical error. Nevertheless, it is remarkable that statistical methods can often achieve usable and credible results in such situations, although they must be interpreted with care. There are two main families of methods for abundance data, and we discuss these next.

2.1 Parametric Abundance Models

In this approach we assume that each species has a certain propensity to enter the sample. (This propensity is not identical to its literal abundance in the population, because the production of the ultimate sample may not transparently represent the underlying population.) We call this the “sampling intensity” of the species: it is the number of representatives of the species expected to enter the sample during

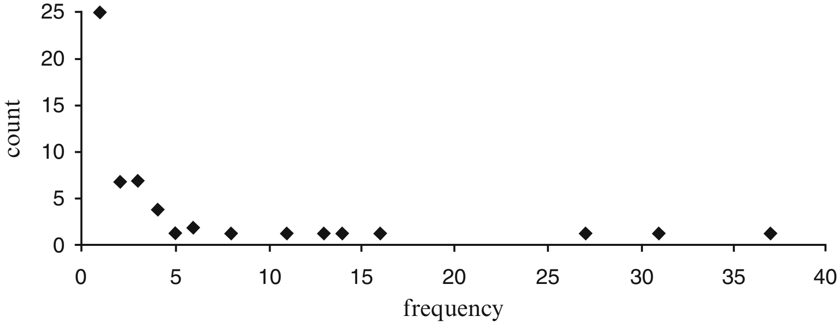

Fig. 1 Frequency count data example

one unit of sampling effort. The simplest possible assumption is that all sampling intensities are equal; however, this almost invariably results in underestimating the true species richness, often severely. It is more realistic to assume that the sampling intensities differ across species, some larger and some smaller. We then model the distribution of the intensities using a probability distribution which in turn is determined by a small number of parameters (for example, its mean and variance). This is a *parametric stochastic abundance model.*

If we suppose that each species independently contributes a random number of representatives to the sample according to the Poisson distribution, with mean equal to the species' sampling intensity, we have a *mixed-Poisson model* for the frequency counts (the "mixing" distribution is the distribution of the sampling intensities). We then fit the mixed-Poisson model to the observed frequency counts, generally via the method of maximum likelihood. (For an outline of the mathematical theory see Chao and Bunge 2002 and Bunge and Barger 2008). This amounts to fitting a "curve" to the data; the curve is projected upward and to the left to obtain an estimate of f_0, the number of unobserved species (i.e., observed zero times in the sample), and f_0 is added to the observed number of species to obtain a final estimate of the total richness. The same mathematical structure yields a standard error (SE) for the estimate, goodness-of-fit assessments, and so forth. Figure 2 shows such a model fitted to our example data. Here the estimate of f_0 is 67, for a total richness estimate of 54 + 67 = 121 species; the associated SE is 39, and the model fit is excellent. (Note that the curve is only fitted to the data up to frequency = 16; we discuss this later.)

Several questions arise immediately. First, how do we choose the parametric model, or stochastic abundance distribution, to use? It would be ideal if basic ecological theory would provide such a model, and it could be confirmed to fit data in a large number of cases. However, while there has been considerable work both in the mathematical and biological literature on the derivation of such models, no consensus has emerged (Williamson and Gaston 2005). Furthermore, it is not clear that a single

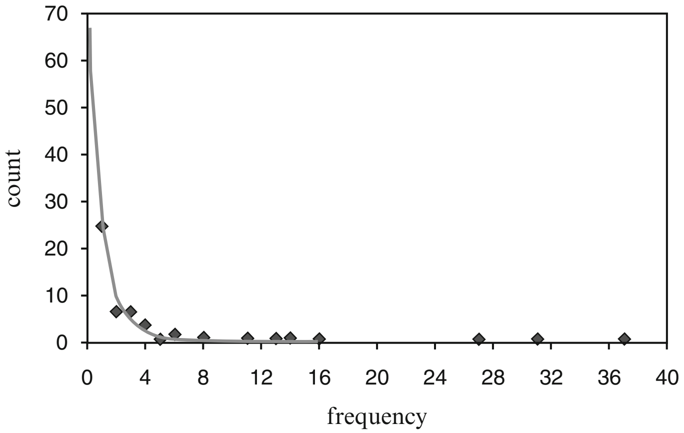

Fig. 2 Frequency count data example with fitted parametric model

model would apply across different types of organisms, or at different levels of the taxonomic hierarchy. Even if such a model could be found for the abundances of microbial species in a specific habitat, under specific conditions, it is not obvious that the model would remain unchanged under the various sampling, molecular biology and bioinformatic procedures that are used to construct a clone library, which is often the final source of data for analysis. One can envision a seamless theory that would mathematically describe environmental abundances, sampling, and the production of the clone library, resulting in a final model for the library as a representation of the underlying population; however, no such theory has yet been attempted. Hence at present the choice of model must generally be empirical, based on apparent fit to the data. We return to this question below.

Second, we typically fit models only to the observed frequency count data up to some maximum frequency, which is called the "right truncation point" or "tuning parameter," denoted by τ. Given an analysis based on the truncated data, we simply add the number of species with frequencies greater than τ to obtain the final richness estimate. (In the example above, $\tau = 16$.) The reason for this is that it may be impossible to find a parametric model (from any known family) that fits a given complete frequency-count dataset. Some researchers justify this truncation heuristically by noting that much of the information relevant to estimation of f_0 may be found in the lower frequencies (more rarely observed species), since these strongly influence the trajectory of the fitted curve, upward and to the left. However, the statistical foundation of this statement is uncertain, and in fact the richness estimate (and its SE) will typically vary with τ.

Nevertheless it is usually necessary to select a value for τ, and there are several possible approaches to this choice. First, we can look for an apparent gap in the data, and use that to subdivide it into "rare" (low-frequency) and "abundant" (high-frequency) species. In the example, such a gap occurs between 16 and 27, so $\tau = 16$ is not unreasonable. This approach requires expert judgment in each case, and furthermore there may be no such gap, or there may be many. Second, we can set τ at a heuristically-selected fixed value, say $\tau = 10$. But if τ can be taken much higher than 10 while still obtaining a good fit, potentially important data points will be omitted unnecessarily. Third, we can set τ to be the maximum value that will allow a good fit of a given parametric model; however, this value (and hence the results of the analysis) will vary depending on the model. Finally, we can take τ to be the maximum observed frequency in the data, and then select the model that gives the best fit to the entire dataset – although no (available) model may be entirely satisfactory. This simple approach tends to be conservative; that is, it tends to produce somewhat lower richness estimates.

As the above considerations show, the choice of model and tuning parameter interact, in a manner that does not currently have a direct statistical resolution. For a fixed value of τ, we can select a model using standard statistical criteria such as the Akaike Information Criterion (AIC). Given a variety of models, each fitted at several possible values of τ, one can examine the Pearson chi-square goodness-of-fit statistic, but at present the final choice of analysis must to some extent rely on expert judgment, pending further research. Fortunately, expansion of the set of available models (based on mathematical research) generally allows τ to increase, as the models become more flexible (Bunge and Barger 2008).

In summary, the advantages of the parametric stochastic abundance model approach are ease of interpretation, clear visual representation, and often excellent fit to a large proportion of the frequency count data. The disadvantages are: empirical selection of tuning parameter and model.

2.2 Nonparametric Abundance Models

In this approach, the basic structure is the same as above: each species is assumed to independently contribute a Poisson-distributed number of representatives to the sample, and the species' sampling intensities are assumed to follow some underlying "abundance" or "latent" distribution. However, here this distribution is not restricted to be a member of a parametric family (such as the lognormal), but rather is allowed to range over the entire set of possible distributions. Rather than fitting a parametric maximum likelihood estimate of total richness, this method computes a nonparametric maximum likelihood estimate or NPMLE (Wang and Lindsay 2005). There are several differences from the parametric approach.

- The final fitted abundance (sampling intensity) distribution is constructed by the procedure for each dataset, rather than being selected from a parametric family. (In fact the nonparametrically fitted distribution is discrete, essentially defining a small number of categories of species abundances.) This has the advantage of great flexibility, but gives no indication as to the suitability of well-known parametric models; essentially it builds a model de novo for each analysis.
- Since the method selects, or rather constructs an approximation to, a single model among the entire class of possible models (not just within a specific parametric family), the question of model selection among different parametric families does not arise.
- The NPMLE is relatively insensitive to the truncation point τ, so the question of τ, and its problematic interaction with model selection, does not arise, or is less important.
- There is not (at present) a direct variance (SE) formula for the NPMLE; instead the SE must be computed by some form of resampling such as the bootstrap. However, the bootstrap must be carried out very carefully, because the species richness estimation problem exhibits certain pathologies which may render bootstrap variance estimates inaccurate.
- There are several possible NPMLE's. All are based on the Poisson sampling model; the differences arise because of the various ways of stabilizing (reducing the variance of) the estimators, and in the approximations and algorithms used for computing them. Because the NPMLE approach is fairly recent, not enough experience has accumulated to make a definitive selection among these.
- It has been argued mathematically in a certain nonparametric theoretical framework that only lower bounds for the total richness can be regarded as reasonable (Mao and Lindsay 2007). Intuitively, this is because the nonparametric approach attempts to make minimal assumptions about the population structure (the underlying abundance distribution), and consequently one must always allow for

an arbitrary number of arbitrarily small species. Some versions of the NPMLE incorporate a "penalty parameter" or other devices to reduce this possibility.

As of this writing software to compute the NPMLE (and its variants) is not readily accessible, and therefore we do not compute the estimate for our example data. However, we expect software for this approach to enter the mainstream in the near future.

2.3 Coverage-Based Estimation

The *coverage* of the sample is the proportion of the population represented by the species that have been observed. For example, suppose there are four species in the population, say A, B, C, D, with proportional abundances 75%, 20%, 4% and 1%, respectively. If species A and C (only) appear in the sample then the coverage is 79%. (There is a statistical subtlety here: the *actual* coverage is random and depends on the particular sample observed, while the *expected* coverage is the long-term average coverage produced by the sampling procedure. Some methods address one version, some the other.) Estimation of coverage may be an easier statistical problem than richness estimation, and has its own literature (Mao 2004). The canonical example of a richness estimator based on this concept is the Abundance-based Coverage Estimator or ACE (Chao 2005). This estimator first inflates the observed number of species in inverse proportion to a nonparametric coverage estimate (the smaller the coverage, the greater the inflation), and then further adjusts nonparametrically according to the variability of the observed frequency counts. Many variations of ACE have been studied, including ACE1 for (apparently) higher-diversity populations (Chao and Lee 1992).

ACE and its variants:

- Are nonparametric and hence do not require model selection
- Admit direct variance estimation (SE), without resampling
- Are sensitive to the choice of τ – typically both the richness estimate and its SE increase with τ
- Provide a sequence of related estimators, typically assuming higher degrees of diversity in the population
- Are known to underestimate total richness in high-diversity populations (i.e., are downwardly biased in such cases)
- Can be theoretically (mathematically) related both to certain families of parametric models (such as the gamma-mixed Poisson/negative binomial), and to the NPMLE framework (Chao and Bunge 2002; Mao and Lindsay 2007)

In our example above, the preferred coverage-based estimator at $\tau = 10$ is ACE, which gives an estimate of 86 with SE of 14, and at $\tau = 16$ (the value selected by the parametric procedure) the preferred estimator is ACE1, which gives 152 with SE of 50. In some datasets the possible values of τ may vary by one or more orders of magnitude, with corresponding variation in the coverage-based estimates.

2.4 Discussion

Given the variety of approaches to estimation of species richness, which method should we use in a given case, and how should we select a final analysis to report? There are several principal considerations. First, the method used should correctly represent the researcher's operative assumptions about the underlying (target) population, sampling procedure, and biochemical system for production of the final frequency count data. All of the methods described above share the same fundamental framework: they assume that each species has a given sampling intensity, the intensities vary in such a way that they can be described by a probability distribution, and members of each species enter the sample independently according to their sampling intensities. If the data can be assumed to have been generated according to this framework, then all of the methods above are equally applicable. (We consider different data structures in Sect. 3).

The simplest and (in our view) best approach would be to compute all estimates and associated statistics such as SE's, goodness-of-fit assessments (where applicable), graphical representations, etc., and examine them all, because each method will illuminate a different aspect of the data. In the best-case scenario all methods will agree (approximately), leading to good confidence in the results. If there is strong divergence between the various analyses, this too is informative, and indicates that a more conservative approach should be preferred until further (or auxiliary) data becomes available. However, as of this writing there is no comprehensive and easy-to-use software package to carry out such a multimethod, parallel analysis. The coverage-based methods do not present major programming challenges and have been widely implemented in software. The parametric and nonparametric maximum likelihood methods require nontrivial algorithm and program development, and in some cases entail significant computing time, especially when many analyses are requested simultaneously, e.g., for multiple values of τ. At present, the existing software for these methods is not comprehensive or integrated, and the applied user will still need expert guidance, which may be available from the software authors. However, software for all methods is undergoing rapid development, and some software with reasonable user interfaces is beginning to appear, even for the computationally-intensive methods.

We recommend that the applied researcher make every effort *not* to be limited by easy-to-use, readily available software. Instead, we recommend that the researcher seek expert advice, which is available in most institutions from a statistics department or group, and analyze his or her data using as many of the above methods as is feasible, and in particular to compare the results of more than one parametric model (for example, the gamma-mixed Poisson or negative binomial model is usually too inflexible to accommodate the high-diversity data often encountered in microbial diversity research). The combined judgment of the biological and statistical experts can then be brought to bear on the results of the analyses. Such an approach will yield a range of results under slightly different model assumptions, and will tend to guard against over-optimistic acceptance of any single result, which may be biased downward or upward, unbeknown to the researcher.

Table 1 Comparison of abundance-based methods

Method	Advantages	Disadvantages
Parametric maximum likelihood	Well-understood properties; represents data via smooth distribution; responds stably to τ; natural visualization; tests suitability of specific abundance distributions	Model selection not obvious; results depend on (and model selection interacts with) τ; computationally intensive; patchy implementation in existing software
Nonparametric maximum likelihood	Does not require model selection; robust across a wide range of abundance distributions; apparently insensitive to τ	Procedure not yet thoroughly studied; standard error must be obtained by resampling; computationally intensive, and software not yet readily available
Nonparametric coverage-based	Well-understood properties; does not require model selection; robust across a range of abundance distributions; several user-friendly software implementations	Tends to be biased downward in high-diversity situations; sensitive to τ; little diagnostic information for choice of τ and specific estimator; no graphical representation

Table 1 summarizes the salient pros and cons of the various methods at the present time.

We emphasize that while this is the current state of the art, progress is rapid and we expect many of the disadvantages listed above (for all methods) to be ameliorated in the near future, especially in terms of computation.

3 Incidence Data

In this scenario we collect species occurrence or incidence data on several different sampling "occasions," or from several different lists; this is also known as capture–recapture, multiple recapture, or multiple list data. For example, the Table 2 shows 10 different samples or lists, which yielded a combined total of 15 observed species. (This is a subset of a larger dataset with 46 samples and 3,717 observed species, extracted from GenBank at the 90% similarity level (Epstein and Bunge 2008).) Each row represents the "capture history" of a particular species (arbitrary species ID numbers are assigned), and each column represents the list of species observed on a given sampling occasion. Note that on each occasion, the only presence or absence – the "incidence" – of each species is recorded, where 1 indicates that the given species was observed on the given occasion (0 otherwise). The right-most column gives the total number of observations for each (observed) species. Analysis of such data has a long history and immense literature, dating back at least as far as the eighteenth century (Borchers et al. 2002, Chao and Huggins 2005); here we attempt only a sketch from a particular point of view.

Table 2 Example multiple recapture data

	Sample										
Species ID	1	2	3	4	5	6	7	8	9	10	Total
1	0	1	0	0	0	0	0	0	1	0	2
2	0	0	0	0	0	0	0	0	1	0	1
3	1	1	1	0	0	0	0	0	1	0	4
4	0	0	1	0	0	0	0	0	0	0	1
5	1	0	0	0	0	0	1	1	1	0	4
6	0	0	0	0	0	1	0	0	0	0	1
7	1	1	1	0	0	1	0	0	1	0	5
8	0	0	1	0	0	0	0	0	1	0	2
9	0	0	0	0	0	1	0	0	1	0	2
10	0	0	0	0	0	1	0	0	0	0	1
11	0	1	0	0	0	0	1	0	1	0	3
12	0	0	0	0	0	0	0	0	0	1	1
13	0	0	0	0	0	0	0	0	1	0	1
14	0	0	0	0	0	0	0	0	1	0	1
15	0	0	0	0	0	0	0	1	0	0	1

We note that in some cases the actual "abundance" or number of observations of a given species on a given occasion may be recorded, leading to integer entries greater than 1; this may be called "multiple abundance data," although there is no standard terminology in this case. Clearly multiple incidence data can be derived from multiple abundance data, but not the reverse. Note also that frequency-count data can be derived from the marginal totals, but the full table cannot be recovered from the frequency-count data. In this example, the latter is (1,8), (2,3), (3,1), (4,2), (5,1). Here the maximum possible frequency is equal to the number of samples. It is possible to apply frequency-count methods to such data if the number of samples is large enough; we return to this issue below.

Multiple incidence data is more highly structured than frequency-count data, and its statistical analysis admits more variations; here we only attempt an outline of the areas we see as especially relevant to microbial ecology. Much of the literature in this area originated with population size estimation for macro-fauna – birds, mammals, fish – and in this application a row in the table represents the capture history for a particular animal. A certain taxonomy of models has emerged from this literature (Borchers et al. 2002). It is not ideally adapted to microbial ecology applications but it has become a de facto standard, and (at least) the first four models are noteworthy here.

- M$_0$: Global homogeneity. Each species is equally likely to occur, on each occasion, i.e., each species has the same sampling intensity or abundance, and the "sampling effort" is the same on each occasion.
- M$_t$: "Time" (t) effect. All species have the same sampling intensity or abundance, but sampling effort varies with occasion or time. Thus all species have the same probability of occurrence on a given occasion (within a given column of the table), but this probability varies across occasions (columns).
- M$_h$: Heterogeneity (h) effect. The species have different (heterogeneous) sampling intensities or abundances, but sampling effort is the same on each occasion.

Thus a given species' probability of occurrence is the same on every sampling occasion: the probability of occurrence is the same within a given row of the table, but varies from row to row.

- M_{th}: Time and heterogeneity effects. The species have different sampling intensities, *and* sampling effort varies across occasions. Thus the occurrence probabilities vary both across rows and across columns.

Further models involving behavioral (*b*) effects have been studied in the literature. These were originally intended to account for individual animals' responses to being captured (e.g., becoming "trap-happy" or "trap-shy"), and while the statistical models for such effects may have an alternative interpretation in the microbial ecology setting, we do not discuss them here. In the following discussion we assume that the sampling occasions or lists are (statistically) independent. There are many situations in which this assumption may be false, but we will retain it here for simplicity, and because it is often reasonable in microbial ecology.

The models M_0 and M_t, which assume equal species abundances, admit straightforward maximum likelihood estimates and can be dealt with in a reasonably uncomplicated manner, statistically and computationally. Unfortunately, the equal-abundance assumption is rarely if ever realistic, and causes severe downward bias when the true population is heterogeneous. Thus the models of interest here are M_h and M_{th}. The more general of these is of course M_{th}; however, it may possible for the researcher to specify, on substantive grounds, whether the "time" effect should be assumed to be present or not, that is, whether sampling effort can be assumed to be constant across occasions. Thus both M_h and M_{th} are potentially useful.

The same three classes of estimators – parametric maximum likelihood, nonparametric coverage-based, and nonparametric maximum likelihood – exist for incidence data as well as abundance data. If we restrict our attention to models M_h and M_{th}, this gives six potential families of statistical procedures to consider. We will briefly consider each of the six possibilities. In addition, there is (at least one) alternative approach based on "estimating equations," which has both parametric and nonparametric aspects, and admits an elegant extension even to the most general model M_{bth} (Chao et al. 2001). However, in its present form this method depends on the "time" order of the samples, i.e., it is not invariant to permutation of the lists. Since this assumption appears to be more adapted to certain kinds of animal-trapping surveys, we do not delve into it in detail here.

3.1 Parametric Incidence Models

3.1.1 Model M_h

In this case, the row totals of the data, that is, the number of times each species is observed, are binomial random variables, where the binomial "success probability" (the probability that a given species is observed on a given occasion) varies from row to row. A parametric mixture model, analogous (and in some ways equivalent)