Estimating the Number of Species in Microbial Diversity Studies

John Bunge,[1] Amy Willis,[1] and Fiona Walsh[2]

[1]Department of Statistical Science, Cornell University, Ithaca, New York 14853; email: jab18@cornell.edu, adw96@cornell.edu

[2]Federal Department of Economic Affairs, Education and Research EAER, Research Station Agroscope Changins-Wädenswil ACW, Bacteriology, 8820 Wädenswil, Switzerland; email: fiona.walsh@agroscope.admin.ch

Annu. Rev. Stat. Appl. 2014. 1:427–45

First published online as a Review in Advance on August 30, 2013

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

This article's doi: 10.1146/annurev-statistics-022513-115654

Keywords

species richness, sample coverage, number of classes, α-diversity, mixed Poisson, zero truncation

Abstract

For decades, statisticians have studied the species problem: how to estimate the total number of species, observed plus unobserved, in a population. This problem dates at least as far back as 1943, to a paper by R.A. Fisher. These methods have found many applications in general ecology, but their importance has grown considerably in recent years, driven by the introduction of high-throughput DNA sequencing into microbial ecology. We examine the state of the art in terms of estimating the total number of taxa in a microbial population from a sample of sequences. We focus mainly on estimating the number of species within a single population (α-diversity), but we also briefly consider statistical inference for comparing the numbers of species across populations (β-diversity). We discuss the full range of statistical techniques, parametric and nonparametric as well as frequentist and Bayesian, and specific implications of their use in microbial diversity studies. We conclude with some recommendations for theoretical investigation and computational tool development.

1. INTRODUCTION

1.1. The Species Problem

Since Fisher et al.'s (1943) original paper, an avenue of research in statistics has dealt with estimating the number of species in a population. A 1993 review of the topic considered more than 500 papers and reviewed 120 (Bunge & Fitzpatrick 1993). Since then, approximately 150 new methodological papers have appeared in the field of statistics, with 100 times that number if we include applied papers from ecology, computational biology, and other interested fields. Existing research runs the gamut of statistical approaches, from parametric to nonparametric, frequentist to Bayesian, and highly theoretical to purely data analytic. The methods developed have found application not only in ecology and biology but also in many other fields in which researchers must estimate the number of classes in a population: computer science, linguistics, astronomy, the social sciences, and even numismatics. By any standard, the species estimation problem is relevant and continues to be explored by theoreticians and practitioners alike.

Over time, researchers have conducted many scientific studies to estimate the taxonomic diversity of a given population, and statisticians have collaborated with biologists to solve the immediate applied problem and to investigate the performance of their methods in the real world. However, because the applications of the diversity estimation problem are so varied and the methods applied so different, a standard practice has not yet come to the fore. This delay is partly because the field continues to evolve rapidly at the theoretical level, partly because different populations require different approaches, and partly because a local solution to an applied problem has often appeared sufficient. The situation is now beginning to change. The problem of estimating microbial diversity, especially in metagenomic studies based on high-throughput DNA sequencing (HTS), is acting as a great attractor, drawing statistical methods and statisticians to itself. This attraction is driven by a scientific problem of startling magnitude and importance and by the unprecedented availability of massive sequence data sets. Under the impetus of this application, which is generating an exponentially increasing quantity of published research, we are seeing an invigorated debate about statistical methodology and widening collaboration between microbiologists and statisticians.

1.2. The Size of the Microbial World

The microbial world is immense, in both the physical and scientific senses. There are $\sim 10^{30}$ prokaryotes (bacteria and cyanobacteria) in the Earth's biosphere (Whitman et al. 1998), and as Foster et al. (2012) have stated, "We live in a microbial world, with microscopic organisms filling discrete ecosystems in such environments as soil, lakes and oceans, the human gut or skin, and even computer keyboards" (p. 420). In particular, bacteria make up most of the Earth's biomass (Whitman et al. 1998), and 90% of the cells in a human body are bacterial (Sears 2005). Vast scientific effort has been and is being devoted to characterizing the importance and function of the microbiome, both locally (e.g., at a single site in the human body) and, literally, globally. Indeed, most authors of scientific papers on this topic preface their remarks with an appreciation of the magnitude of the microbiome.

The microbial world is also enormous from the perspective of the current limitations of human knowledge. The principal problem is arguably uncultivability, or the great plate count anomaly: The vast majority of bacteria cannot be cultured in the laboratory. In fact, "studies of environmental 16S rDNA showed that the anomaly is more than a gap in the total cell count—more than 99% of all *species* from various environments are 'unculturable'" (Lewis 2009, p. 182; see also Amann et al. 2001, Staley & Konopka 1985). Unculturable microorganisms can be found in nearly every group within the Bacteria and Archaea, and several groups at the division level have been identified with

"no known cultivable representatives" (Lewis 2009, p. 182). Until recently, then, scientists have known that the diversity of microbial life is very great but have had difficulty gaining access to this diversity; the irreproducibility of the natural world appears to be a very challenging problem. According to Gilbert et al. (2011), "the vast imbalance between what it is possible to hypothesize and test, and what is unknown means that every microbial ecologist is on an epic voyage of equal importance to that of Darwin" (p. 2).

1.3. The Sequence Data Deluge

Like many fields, the study of microbial diversity has undergone a big-data revolution in recent years. This change mainly stems from the development of HTS, sometimes called next-generation or massively parallel sequencing. The most important differences between HTS and traditional Sanger sequencing are throughput and the identification of previously unknown DNA sequences. Logares et al. (2012) noted that "while a typical Sanger run would generate 10^2 sequences (600–900 bp of length), HTS (e.g., 454 and Illumina) can potentially generate 10^6–10^9 sequences (100–700 bp) per run" (p. 107). HTS has enormously improved the exploratory capacity of metagenomic studies. As Gilbert et al. (2011) wrote, "In metagenomics, we isolate DNA directly from the environment and use it to characterize the taxonomy and function of the biological community in that ecosystem. . . . Our ability to interpret these data is always improving . . . and we stand on a precipice of unprecedented discovery" (p. 2).

For the purpose of estimating the number of microbial species, the procedure is roughly as follows. First, DNA sequences are extracted from a microbial sample (Foster et al. 2012, Logares et al. 2012). Apart from the applied science and technology involved, this is a nontrivial process with several pitfalls, which we mention briefly in Section 2.1. Next, because no unanimous concept of distinct species exists for microbial life, decision rules for grouping the observed sequences into classes are also contested. Microbial ecologists use operational taxonomic units (OTUs), formed by grouping similar sequences together. Typically, sequences sharing 97% identity are grouped into the same OTU, although any level of similarity can be used, and indeed multiple levels may be analyzed in the same study. OTUs are formed using clustering algorithms, whose performance is often not fully understood, and results are sensitive to the algorithm used. Finally, the numbers of members in the OTUs are counted. These counts become the abundance or frequency count data, which are then subjected to statistical analysis. The frequency counts are f_1, the number of OTUs having one member, or the singletons, which play an important role; f_2, the number having two members, or the doubletons; f_3, the number having three members; and so on. The objective of statistical analysis is to estimate (technically, to predict) the number of unobserved species, f_0, and hence the total number of species, unobserved plus observed: $f_0 + f_1 + f_2 + f_3 + \cdots$.

Figure 1 shows two such data sets, both from the same study, on the effects of the use of antibiotics on bacterial populations in soil ecosystems (F. Walsh, S. Owens, B. Duffy, D.P. Smith, J.E. Frey, submitted). (The first data set was subsequently revised slightly, but the original version suffices for our demonstrations here.) The samples were isolated from two different soils. Data set S3a has 11,338 sequences grouped into 1,187 OTUs, and data set 321 has 13,907 sequences in 1,005 OTUs. These numbers are not large by HTS standards, but they make good examples for our discussion, and they display all of the typical features of microbial diversity data. Note the typical (mirrored) J shape of the curves, indicating large numbers of rarely observed species: 317 and 277 singletons for S3a and 321, respectively, followed by a long slow decay to the right denoting a few very abundant species, especially in data set 321. Researchers are rapidly increasing the size of such data sets: The authors recently received one from a single study having 12.6×10^{10}

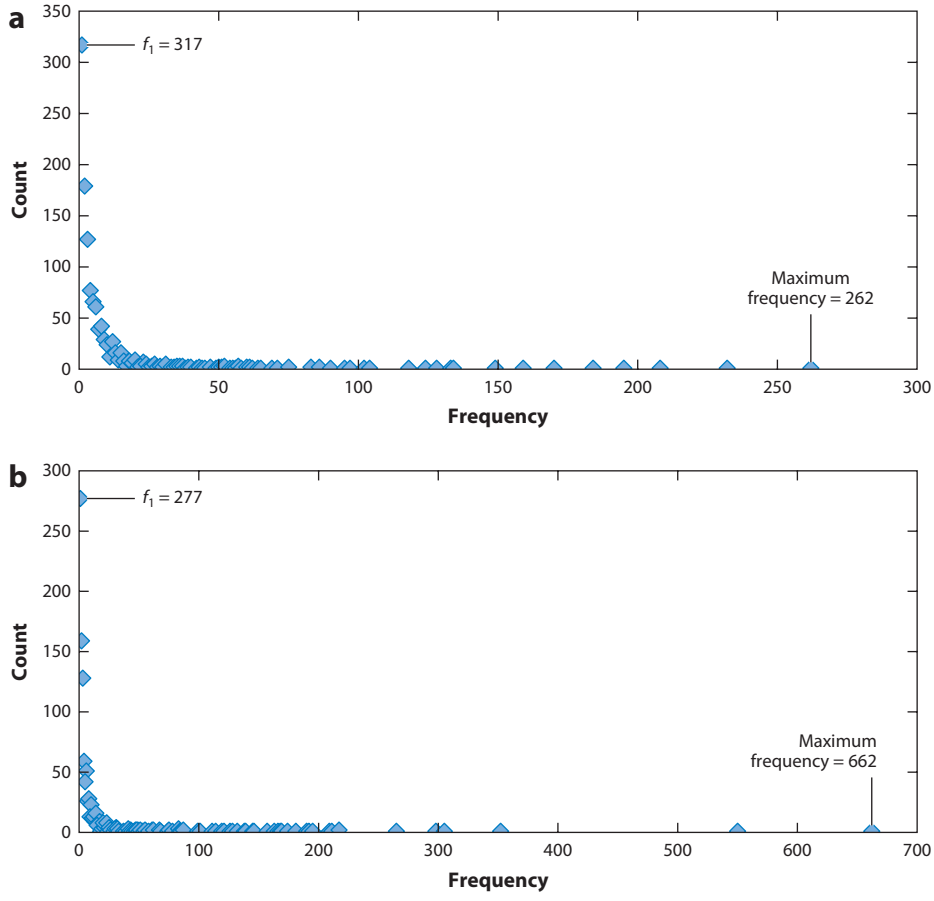

Figure 1

Complete frequency count data set (*a*) S3a and (*b*) 321 shown with frequency as a function of count. Singleton count, f_1, and maximum frequency are indicated.

sequences and 11.5×10^{10} OTUs—and 10 billion singletons—which exceeded certain maximum allowed values (Int32) in our software CatchAll (Bunge et al. 2012b; see Section 3.3).

1.4. What Is Diversity?

Our task, then, is to estimate the total diversity of the microbial population under study—observed plus unobserved. We should first clarify just what we mean by diversity. Ecologists distinguish among three aspects of diversity, called α, β, and γ (Tuomisto 2011). The first is the diversity within a single population or system, the second is comparative across two or more populations, and the third refers to a global measure across multiple populations. In this review, we concentrate on α-diversity, with a brief excursion into β-diversity in Section 6 (we do not go into γ-diversity).

Leaving aside qualitative concepts, several candidate quantitative definitions or indices of α-diversity exist. Let S denote the number of species in the population, and suppose that the species occur in proportions $p_1, \ldots, p_S$. Various formulae have been studied: For example, Shannon's entropy index is $-\sum_{i=1}^{S} p_i \ln p_i$, and Simpson's index is $\sum_{i=1}^{S} p_i^2$. These and other such

indices combine the total taxonomic richness S with some measure of evenness of the population. Such quantities allow various interpretations and are statistically estimable in many settings (see Valiant & Valiant 2011 and Zhang 2012 for entropy and see Zhang & Zhou 2010 for Simpson's index). However, their physical interpretation is not transparent, nor is it obvious how to directly compare values across populations.

Here we focus on the simplest possible diversity index, the number of species, S. We assume that S is fixed (i.e., not variable within a given population or in the course of a given study) and finite (i.e., we exclude scenarios in which S may vary or increase without bound) but unknown. S is often called the species richness (Tuomisto 2011), although the latter term is sometimes used to mean diversity more broadly defined, as in the preceding paragraph. The number of species is readily interpretable and comparable across populations but more difficult to estimate because it counts every species, no matter how small. Because of the latter point, some statisticians advocate searching for only a (greatest) lower bound for S rather than a point estimate, and in a fully nonparametric setting, researchers have shown that only lower bounds are possible (Mao & Lindsay 2007). We return to these matters in Section 3.4.

We conclude this section with two observations. First, statistical methods for estimating the number of taxa or other quantities such as entropy have a wide range of applications beyond microbial ecology. (Methods for general ecology of nonmicrobial organisms largely overlap with those presented here, so we do not discuss them separately.) Outside of biology, the main application is arguably in computer science and computational linguistics. Ohannessian & Dahleh (2012) wrote, "On a pragmatic level, any engineered system, be it for inference, communication, or encryption, requires working with a finite number of symbols. Therefore, the most straight-forward model is a finite alphabet" (p. 1). Here, the characters of the alphabet correspond to the taxa. In this context, statistical analysis of various kinds of finite sets arises in a number of ways, including estimation of the size of an alphabet or vocabulary, of the information in a signal train, and of database complexity. For recent entrées to the large literature of this area, we refer the reader to Bhat & Sproat (2009), Ohannessian & Dahleh (2012), and Valiant & Valiant (2011), and references therein.

Second, the species problem is tangent to the area of capture-recapture statistics (Bunge 2013). The fundamental difference is the granularity of the data structure: In the species problem, we have frequency count or abundance data as described above, whereas capture-recapture analysis is based on incidence data. In the latter scenario, there are, e.g., k trapping or observation occasions, and on each occasion, a given species (or, more commonly, a uniquely identifiable member of the population) may or may not appear in the sample. The incidence matrix then has a row for each observed species (or individual) and a column for each trapping occasion, and we enter a 1 if the given species (or individual) is observed on the given occasion, 0 otherwise. Converting incidence to abundance data (by summing the columns) is possible, but the reverse is not. In both settings, the basic objective is the same: estimating the number of unobserved species (or individuals) and hence the total number, observed plus unobserved. In this review, we focus exclusively on abundance data, which is (currently) typical in microbial diversity studies, and we refer the reader to the references in Bunge (2013) for capture-recapture.

2. INFERENTIAL ISSUES

2.1. Data Quality

HTS is prone to errors of various types. This problem has an important effect on subsequent statistical analyses. Quince et al. (2011) stated, "The large read numbers obtainable mean that the absolute number of noisy reads is substantial. Consequently it is critical to distinguish true

diversity in the sample from noise introduced by the experimental procedure" (p. 2). For example, sequencing errors can cause two sequences that should be grouped into the same OTU to be classified as distinct OTUs and possibly even as singletons if no match is found. This misclassification leads to artificially inflated low-frequency counts (especially f_1), which in turn cause a disproportionate inflation of statistical diversity estimates. In particular, f_1 plays an important role in many estimates, further exacerbating the problem. The biochemical details of the potential errors are beyond the scope explored here, but researchers have devoted great effort to correcting the errors at the bioinformatics-processing stage, with considerable success, leading to better but still not perfect data quality (Logares et al. 2012, Quince et al. 2011). In some settings, e.g., the analysis of viruses, spurious low-frequency counts remain problematic, and this difficulty has led statisticians to try to compensate ex post facto, statistically, for uncertainty in the low-frequency counts. From one perspective, this compensation is analogous to modeling that accounts for measurement error in the observed variables. We return to this topic in Section 3.6.

2.2. Sample Versus Population and the Goal of Inference

In any statistical analysis, making a clear distinction between the sample, which provides the data under analysis, and the population or universe from which the sample was derived is crucial. Surprisingly, though, this seemingly natural distinction is not entirely clear in many microbial ecology contexts. Suppose, for instance, that we collect 1 liter of seawater from a depth of 500 m at a particular site in the ocean. From the 1 liter, we remove 5 ml and subject it to exhaustive analysis, sequencing all available microbial DNA and clustering the resulting sequences into OTUs. Based on the OTUs, we calculate the frequency count data and compute an estimate of total diversity (species richness) using one of the methods discussed in Section 3. The question is, to what population quantity does this diversity estimate refer? That is, the claimed total species richness is the richness of what population? Is it the richness of the 1-liter original sample, a nearby region of the ocean, the entire ocean, or the entire pelagic microbiome? No obvious answer to this question exists in many cases.

Consider the following metaphor or allegory (Böhning & Schön 2005). We go to a freeway interchange and record the numbers of occupants, including the drivers, of the passing cars, for a fixed period of time, e.g., during the daylight hours of a given day. In principle, we could then extract frequency count data: f_1 = the number of cars with one occupant, f_2 = the number with two, and so forth. From this data, we could in principle compute a statistical estimate of f_0, the number of cars with zero occupants, i.e., cars not presently being driven, and hence the total number of cars in existence, but (apart from many other potential objections to this procedure) cars in existence where? That is, what is the target population? Is it all cars that might pass by our interchange, which could be all cars in North America, or just some local subset thereof? The answer is unclear. The reader may object that no one would carry out such a study, but it is not too different from the ocean-sampling example discussed above. In that setting, what is the microbial population under study? All microbial organisms that may enter the 1-liter collection bottle, or some more restricted subpopulation?

Answering these questions will require the combined efforts of microbial ecologists and statisticians, and the answers may well vary from study to study. For the present, we adopt the following definition of the population under consideration: It is what would be observed if the operative sampling and analysis protocols were carried out to infinite effort. This definition is at least a plausible thought experiment, although it may or may not correspond literally to the population that the scientist wishes to analyze.

3. THE STANDARD STATISTICAL MODEL

3.1. The General Model

Having explored the implications of generating such an estimate, we turn to the statistical question of estimating diversity. Two main approaches to modeling the species-sampling or data-production processes exist. In the first, we assume that each species independently contributes a Poisson-distributed number of representatives to the sample but at different rates (i.e., with different means), $\lambda_1, \lambda_2, \ldots, \lambda_S$. Here, λ_i is the number of members of species i expected to appear in the sample after one unit of sampling effort. In the second, we assume that the species occur in the population in proportions $p_1, p_2, \ldots, p_S$ ($p_1 + p_2 + \cdots + p_S = 1$) and that sampling is multinomial, so that the expected number of representatives of species i in a total sample of size n is np_i. In the Poisson model, the number of distinct individuals (as opposed to species)—here, sequences—is random, and in the multinomial model, it is fixed. The former can be converted into the latter by conditioning on the sample size, but for various reasons, the Poisson model is more mathematically tractable, so we focus on it here.

Let X_i denote the number of members of species i that appears in the sample, so that X_i is a Poisson random variable with mean λ_i. X_i is observable only if $X_i > 0$. Estimating all of the λ_i individually is impossible, so we make the simplifying assumption that the λ_i themselves are a random sample from some distribution, called a mixture or sometimes a species abundance distribution, although it is really the distribution of sampling intensities, which may or may not be directly related to the literal abundances of the species in the population. Then the observable random variables are the strictly positive X_i, and they follow a zero-truncated mixed Poisson distribution. (Equivalent mixing distributions exist so that the mixing can be done before or after zero truncation; see Mao & Lindsay 2007, Valero et al. 2010.) Thus, the observed frequency counts are $f_1 = \#\{X_i: X_i = 1\}, f_2 = \#\{X_i: X_i = 2\}$, and so on. The zero count, f_0, is the (random) number of unobserved species, and our goal is to estimate (again, technically, predict) this quantity.

This model is the basis for almost all existing research on the species problem. We may, though, point out two restrictive aspects of it: It has structural independence assumptions (for example, the λ_i are independently sampled from some underlying distribution), and the marginal distribution of each species's sample count must be mixed Poisson. These assumptions have consequences (discussed in the following sections), which can lead to difficulties in fitting the standard model described above to microbial diversity data. Hence, extending or generalizing the standard model is reasonable, and we return to this idea in Section 4.

3.2. Data-Analytic Considerations

Species abundance samples in microbial ecology almost invariably display two salient characteristics. First, there is a steep slope upward to the left, indicating many low-frequency observations. In some cases, the number of singletons may be 10 or more times the number of doubletons. Second, there is a long and sparse tail to the right, indicating the presence of a few very high-frequency observations, i.e., only a few highly abundant (dominant) species. In some cases, the right-tail frequencies may jump by orders of magnitude: 100, 1,000, and 10,000. The left slope at least tends to be smooth and can be accommodated in various ways, such as by taking the abundance distribution to be itself a finite mixture of distributions with different slopes. The right tail is more problematic. Any mixed Poisson distribution will have a smooth right tail, and hence large outliers at wide intervals will challenge goodness-of-fit statistics.

Various solutions have been proposed. The most common is to simply cut off the frequency count data at some maximum frequency τ, discarding from the statistical analysis frequency counts

above τ and adding them later to the final total richness estimate. This method is equivalent to deleting outliers in, e.g., regression, although in many statistical settings, one can distinguish identifying characteristics of outlying observations, whereas here they are homogeneous with the rest of the data. The question of a suitable choice of τ then arises, and no straightforward answer exists. Some simple estimators (see Section 3.4) use only the lowest frequencies so that τ is necessarily set at 2, for instance. In other cases, $\tau = 10$ has been proposed based on empirical experience and the idea that most of the predictive information in the data resides in the low-frequency counts; this value of τ is encoded in some software (Chao & Shen 2003–2005). Alternatively, one can look for the maximum τ for which a given model displays acceptable goodness of fit; this technique entails fitting a model (or models) at every value of τ and making comparative goodness-of-fit assessments, in turn raising issues of simultaneous inference (multiple hypothesis tests) and potential overfitting (Bunge et al. 2012b). Finally, one can simply set τ to be the maximum observed frequency, i.e., use all of the frequency data in the analysis. This method is appealing, but, in addition to the potential lack of model fit, large outlying frequencies may also cause erratic behavior of some estimators (see Section 3.4). In fact, the problem may partly result from the use of mixed Poisson models, and we return to this idea in Section 4.

3.3. Parametric Frequentist Procedures

The oldest and longest-studied approach to the species problem, dating back to Fisher et al.'s (1943) paper, assumes that the species abundance distribution, i.e., the mixing Poisson distribution, is parametric, with a low-dimensional parameter vector. More technically, we regard the sampling intensities of the species, $\lambda_1, \lambda_2, \ldots, \lambda_S$, as the realizations of random variables $\Lambda_1, \Lambda_2, \ldots, \Lambda_S$, which are independent and identically distributed (i.i.d.) according to F. F is indexed by a parameter vector θ, which in applications typically has one to seven parameters. In principle, F can be any parametric distribution, discrete or continuous, or even concentrated at a single point (in the latter case, all species have the same sampling intensity or abundance). To take a simple continuous example, F may be the exponential distribution, which has one parameter (the dimension of θ is 1); in this case, the F-mixed Poisson is geometric.

Let $p_\theta(0)$ denote the resulting mixed Poisson probability of zero, i.e., $p_\theta(0) = E_\theta(e^{-\Lambda}) = \int e^{-\lambda} dF(\lambda; \theta)$, and let s denote the observed number of species, $s = f_1 + f_2 + f_3 + \cdots (s = 1{,}187$ and $1{,}005$ in our two data sets, respectively). If $p_\theta(0)$ were known, $s/(1 - p_\theta(0))$ would be the Horvitz-Thompson estimator of the total richness, S. Instead, we estimate θ by maximum likelihood (ML) from the zero-truncated data, i.e., the frequency counts, $f_1, f_2, f_3, \ldots$, and the resulting empirical estimator $s/(1 - p_{\hat{\theta}}(0))$ is the conditional MLE of S. Standard errors (SEs) are obtained via ML asymptotics, and goodness-of-fit assessments are carried out in the usual way using, e.g., Pearson χ^2 statistics, Akaike information criterion, and Bayesian information criterion (BIC) (Bunge et al. 2012b, Bunge & Barger 2008). Confidence intervals (CIs) are asymmetric, based either on profile likelihood or on a log-transformed approximation due to Chao (1987).

Our software, CatchAll (Bunge et al. 2012b), fits a suite of parametric models, in which F can be a single point mass (the equal-abundance model), a single exponential distribution, or (finite) mixtures of two, three, or four exponentials. For our two examples, CatchAll finds the best F among its available models to be a mixture of three exponential distributions. For S3a, the estimate of S is $\hat{S} = 1{,}824$ (SE = 122, $\tau = 184$), with 95% CI (1,625; 2,112) and a Pearson χ^2 p-value of 0.604, indicating a good fit. For 321, the corresponding results are $\hat{S} = 1{,}482$ (SE = 59, $\tau = 163$), CI (1,380; 1,612), and $p = 0.014$, indicating a questionable model fit. **Figure 2** shows the fitted curves.

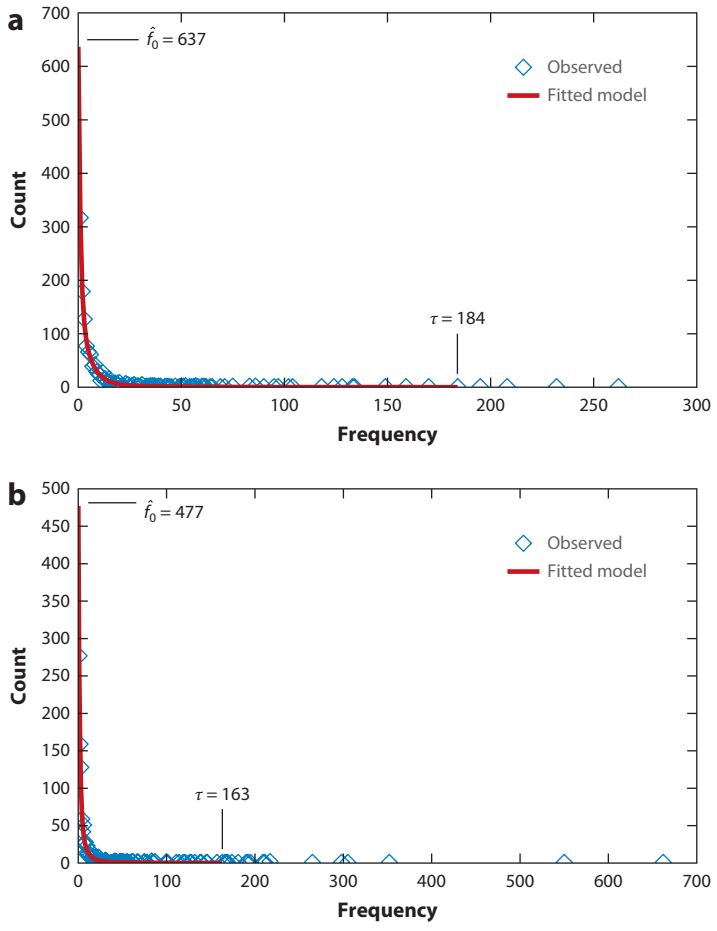

Figure 2

Frequency count data set (*a*) S3a and (*b*) 321 with fitted parametric model. Fitted species abundance model is a mixture of three exponential distributions. Estimated zero count, $\hat{f}_0$, and upper cutoff, τ, are indicated.

The mathematical foundations of ML in this problem were worked out in the 1970s (Bunge & Barger 2008), and computation has improved in the past few years (Bunge et al. 2012b). But the desirable optimality properties of ML depend on the model being correct, at least in principle (the robustness of ML procedures in this problem has been little studied to date). Therefore, apart from the question of the upper frequency cutoff, τ, a major issue remains: the choice of the mixing distribution, F. Two approaches are possible. First, one can look for a theoretical justification. A theory of stochastic speciation processes, for example, might imply a limiting equilibrium distribution that could be used as a basis for estimation. Various distributions have been proposed on this basis, such as the log-normal distribution, but countervailing arguments were quickly produced (Williamson & Gaston 2005). Furthermore, even if such a distribution were theoretically justifiable, there is no guarantee that the actual species abundance distribution in nature corresponds to the sampling intensity distribution that regulates the sampling process; numerous factors beyond the abundances of the species affect the latter, including sampling biases of various kinds and in particular the error-generation process in HTS alluded to above. Hence, the theoretical justification program has had limited success to date.

The alternative is an empirical approach: Try different models F and see what fits. Taking this approach, researchers have applied a wide array of distributions, including the log normal, gamma (and exponential), generalized inverse Gaussian or Sichel (and inverse Gaussian), Pareto, and log-t; we may envision others, such as stable laws (Bunge & Fitzpatrick 1993, Bunge & Barger 2008, Quince et al. 2008). Recently, finite mixtures have come to the fore. Our software, CatchAll, implements a suite of mixtures of exponential distributions with increasing numbers of components, from 0 (i.e., $\lambda \equiv$ a constant) to 4. Wang (2010) proposed finite mixtures of gamma distributions. Such models often provide good fits to data, at least up to some cutoff τ (three or more parameters usually seem to be required). But the purely empirical approach to model selection may leave the analyst with multiple competing data analyses with nearly equivalent fits and possibly divergent results. Thus, the question of parametric model selection remains unsettled.

3.4. Nonparametric Frequentist Procedures

The problem of selecting a species abundance distribution F vanishes if we take a nonparametric approach, i.e., we allow F to vary across the class of all distributions (on the positive half-line) or across some nonparametric subclass thereof (i.e., the subclass cannot be indexed by a finite-dimensional parameter). This approach also has a long history and can be traced back to the work of Alan Turing (Good 1953). Two main approaches exist. The first, due primarily to Chao and coauthors (see Chao 2005 and references therein), is based on the sample coverage, which is the proportion of the population having representatives in the sample (see Section 5). Under the Poisson sampling model, the expected coverage is $1 - P(X_i = 0)$. Note the connection of the expected coverage to $1 - p_\theta(0)$, the denominator of the conditional MLE in the parametric case. From one perspective, the coverage-based estimators can be regarded as incorporating an estimator of the (expected) coverage. The Good-Turing estimator of S is $s/(1 - f_1/n)$ (Good 1953), where n is the number of individuals in the sample—here, the number of sequences—i.e., $n = 1f_1 + 2f_2 + 3f_3 + \cdots$. This estimator is suitable for the case in which all abundances are equal, $\lambda_i \equiv \lambda$, because $E(f_1/n) \approx E(f_1)/E(n) = S\lambda e^{-\lambda}/(S\lambda) = e^{-\lambda} = P(X_i = 0)$ in this case. Chao's abundance-based coverage estimator (ACE) (Chao 2005) is based on the Good-Turing estimator but adds a nonparametric adjustment for heterogeneity, i.e., unequal abundances, which is a multiple of the coefficient of variation of the frequency count data. These estimators can be computed in CatchAll, in Chao's software SPADE (Chao & Shen 2003–2005), and in other packages. In our examples, the ACE results are $\hat{S} = 1{,}444$ (SE $= 30$), CI (1,392; 1,510), for S3a and $\hat{S} = 1{,}245$ (SE $= 30$), CI (1,194; 1,311), for 321. ACE and its variants have some advantages and some disadvantages relative to parametric modeling. One need not select a model F, but no graphical or other goodness-of-fit assessments exist either. The estimates tend to be lower than parametric estimates in high-diversity situations, although a high-diversity variant, ACE1, adds a larger heterogeneity adjustment. The choice of τ is fixed at 10, based both on empirical experience and on the mathematical and empirical fact that the ACE estimators can behave erratically when large outlying frequencies are included in the data. However, the coverage-based estimators are readily computable (although SEs are considerably more complicated) and have been widely popular in microbial ecology.

The second major theme is the nonparametric MLE. Here the idea is to address the mixed Poisson model directly but nonparametrically, allowing an (essentially) arbitrary mixing distribution F (see Böhning & Kuhnert 2006, Böhning & Schön 2005, and Mao & Colwell 2005, along with references therein). It turns out that under this procedure, the final fitted F is a finite mixture of point masses, i.e., a finite number of abundance classes exists, each with a certain probability or weight. This F is comparable to the finite mixture models used in the parametric case, except

there the mixture components are diffuse, not points, and the number of components is selected by the practitioner (either by prior choice or by goodness-of-fit analysis) instead of by the procedure. Numerical algorithms for computing the nonparametric maximum likelihood estimator (NPMLE) are not simple, and other criteria, such as the BIC [rather than the (nonparametric) likelihood itself], may be used to stop the numeric optimization (Böhning & Kuhnert 2006). In addition, SEs, and ultimately CIs, are typically evaluated via a parametric bootstrap based on the fitted model. In our examples, we obtain, using the software CAMCR (Böhning & Kuhnert 2009), NPMLE $\hat{S} = 1{,}461$ (SE $= 172$) for data set S3a, and the fitted F is a four-point mixture. For data set 321, we obtain $\hat{S} = 1{,}210$, again with F a four-point mixture. However, in this case, the bootstrap simulation from the fitted model produces a substantial number of huge outliers that render the SE calculation infeasible and indicate instability in the model. The NPMLE method is promising, but it has not yet been applied in microbial ecology to our knowledge; we expect to see such applications as more empirical experience accumulates and more software becomes available. We refer the reader to Koenker & Mizera (2013) for a current review of the NPMLE for mixture models in a more general setting.

At a more general theoretical level, Mao & Lindsay (2007) carried out a profound analysis of the species problem from a fully nonparametric perspective. They quoted the statistician I.J. Good as saying, "I don't believe it is usually possible to estimate the number of species, but only an approximate lower bound to that number. This is because there is nearly always a good chance that there are a very large number of extremely rare species" (p. 917). Mao & Lindsay found "no locally unbiased and informative estimator . . . , no genuine two-sided intervals and arbitrarily bad informativity when reducing bias to zero. However, . . .there exist nonparametric lower confidence limits" (p. 918). Thus, the researcher is faced with a philosophical dilemma: Parametric methods have many desirable properties but depend on the choice of a specific parametric model, whereas nonparametric methods avoid this choice but cannot achieve the goal of formally estimating the species richness, only bounding it from below. From this perspective, the choice of a method depends on what final empirical claims the researcher wishes, and is willing, to make.

3.5. Bayesian Procedures

Bayesian techniques are especially appealing in the species problem owing to the uncontrolled nature of the unknown, namely the total richness: Establishing a prior distribution for S implies some smoothing on the parameter (and possibly on the entire inference procedure). At present, no literature on nonparametric Bayesian methods in this setting appears to exist (but see Tardella 2002 for the closely related capture-recapture problem), so here we discuss parametric methods only. A Bayesian analysis has two aspects: the overall structure of inference and the particular choices of priors for the various parameters. An appreciable quantity of literature on applying the Bayesian approach to the mixed Poisson model now exists (see Barger & Bunge 2011, and references therein). Most authors take the mixing distribution F to be parametric with parameter θ, and they establish priors separately or independently on θ and S, followed by a standard approach of updating the prior based on the observed frequency count data, yielding a point and interval estimate (e.g., highest posterior density interval) for S.

The spectrum of Bayesian approaches ranges from fully subjective, in which the analyst explicitly incorporates belief or expert opinion into the prior distributions, to formally objective or noninformative, in which mathematical definitions of noninformativeness are used to minimize the information in the priors. Between these poles, one can select priors on an ad hoc basis or select conjugate priors for mathematical convenience. Previous research in this area has typically fallen in this middle ground, but more formal developments recently arose. Quince et al. (2008)

obtained good results in the microbial diversity problem using a variety of mixture distributions F, including the log normal, generalized inverse Gaussian, and log-t. They established a hierarchical Bayes structure with nominally noninformative priors for the parameters of F and an independent flat prior for S. In addition, they used Bayesian model selection procedures to choose a final F. Barger & Bunge (2011) took an explicitly objective Bayes approach. They calculated the global (both θ and S) Jeffreys and reference (objective) priors analytically, finding that the joint prior for θ and S factors into independent priors, thus justifying part of the structure used by previous authors. In addition, they found that the universal reference prior for S is $1/\sqrt{S}$, which can be used regardless of the choice of prior for θ. They obtained good results in analyzing some microbial diversity data sets.

Mathematically, the species richness, S, is an integer parameter and hence requires special treatment, but recent developments have clarified this case (Berger et al. 2012). On one hand, at the applied level, the Bayesian approach has notable advantages, such as the smoothing and model selection mentioned above, the concentration of the posterior distribution on the integers, and the natural asymmetry and easy interpretation of the posterior intervals. On the other hand, the informative parametric approach requires the choice of specific priors for the parameters of F, and the objective approach has nontrivial mathematical complications, especially in the treatment of θ. In both cases, modern Bayesian computation is required, and to date no general software has become available, so applications to microbial ecology are in the process of moving beyond the proof-of-concept stage.

3.6. Estimates Versus Lower Bounds

We discussed above the theoretical justification for considering only lower bounds for the species richness in the strict (frequentist) nonparametric approach. But simpler reasons may exist as well. In many cases, the low-frequency counts, e.g., f_1 and f_2, may be uncertain despite the best efforts at error correction (Section 2.1). In the analysis of viral data, for example, not all of the bioinformatic error-correction procedures may be applicable, or they may not address all sources of error (Allen et al. 2013). For this reason, biologists sometimes specifically request lower bounds for the total richness, even in the parametric setting. Various methods have been proposed to produce lower bounds.

Estimators developed under the assumption of homogeneity, i.e., equal species abundances ($\lambda_i \equiv \lambda$), are well known to be biased significantly downward in the presence of heterogeneity, i.e., unequal species abundances (Böhning & Schön 2005). Hence, estimators for the homogeneous case can be used as lower bounds. These include the Good-Turing estimator mentioned above and the MLE under the homogeneous Poisson model. In our examples, these values are 1,227 (Good-Turing) and 1,187 (homogeneous Poisson) (both at $\tau = \tau_{\max} = 262$) for S3a, and 1,025 (Good-Turing) and 1,005 (Poisson) (again, both at $\tau_{\max}$) for 321. In both cases here, the homogeneous Poisson estimate is equal to the observed number of species, s. A less restrictive estimator, but still a lower bound, is the so-called Chao1 $= s + f_1^2/(2f_2)$, which is 1,468 and 1,246 for S3a and 321, respectively. This bound has been popular in the applied microbial ecology literature. For Chao1, we invariably have $\tau = 2$.

The ideal way to deal with uncertain low-frequency counts would be to correct them at the source, or failing that, to introduce a statistical model for the error process, but the latter method is far from clear at present. Alternatively, one can try to treat the low-frequency counts as potentially erroneous in the statistical estimate of S. In the Bayesian approach, one could apply a prior on S that is either bounded above or rapidly decreasing in some explicit way; we are not aware of any existing research in this regard. In the frequentist approach, researchers have put forth

two ideas for microbial diversity data (Bunge et al. 2012a). First, the data may be regarded as left censored, that is, the frequencies up to some maximum, e.g., 2, may be regarded as indistinguishable so that rather than distinct counts, f_1 and f_2, we have only a single combined count, $f_{1+2} = f_1 + f_2$. Estimation can still be carried out, and the estimate of S is lower than in the uncensored data. However, whether this estimation makes sense scientifically is unclear because this procedure asserts, for example, that a singleton could have been a doubleton (and vice versa), whereas the actual problem seems to be that some of the singletons should not exist at all in the data. Finally, and most drastically, if the mixed Poisson model is used with mixing distribution F, and F is itself a finite mixture (Section 3.3), we can subtract the highest-frequency component (of the finite mixture) from the fitted model and allow the remaining components to determine the estimate of S. We call this approach discounting. Because much of the data is typically concentrated in the low-frequency counts in microbial studies, the discounting method can severely modify the estimate, often resulting in a decrease in the estimate of S by an order of magnitude or more. In our examples, the fitted models F are mixtures of three exponentials for both data sets, so subtracting the highest-diversity component leaves a mixture of two exponentials, and the resulting discounted estimate computed by CatchAll is 1,061 for S3a; CatchAll could not compute the discounted estimate for 321 because the model is too unstable. The discounted estimate 1,061 is less than 1,187, the observed number of species in the sample (S3a), but this estimate is not unreasonable because there were 317 singletons, and we are claiming that many of these (as determined by the proportion assigned to the highest-diversity component of the fitted model) were erroneous or artifactual. In the end, however, none of these methods is wholly satisfactory when compared with correcting errors at the source.

4. NOVEL APPROACHES

4.1. Ratio-Based Methods

Although many of the estimators described above produce apparently sensible results and are based on mathematically powerful techniques, one may ask whether a simpler approach might better reflect the true data-generating process. Conceptually, such an approach would postulate a structure that reflects the manner in which the frequency counts were drawn rather than attempt post hoc to fit a complex model (i.e., the mixed Poisson) that may challenge interpretation. Common distribution structures employed to model count data include the Poisson, binomial, and negative binomial (gamma-mixed Poisson) distributions. In 1945, Katz (see Johnson et al. 2005) showed that these are the only true probability distributions that can occur when the recursive frequency ratio $(j+1)f_{j+1}/f_j$ follows a linear structure, i.e., $(j+1)f_{j+1}/f_j = b_0 + b_1 j$. When this scaled ratio is plotted as a function of j, many microbial diversity data sets do, in fact, display an approximately linear structure. This linearity can be used for diversity estimation because if a linear regression model can capture the structure of the frequency ratios (at least the contiguous ones, i.e., those j for which $f_j \neq 0$), then the estimated regression coefficients and SEs can be employed to give estimates (predictions) and confidence (prediction) intervals for the number of unobserved frequencies, i.e., f_0. Rocchetti et al. (2011) used a weighted log-transformed regression model to capture the shape and heteroscedasticity of $(j+1)f_{j+1}/f_j$. In our examples, this model yields $\hat{S} = 2{,}286$ (SE $= 573$, $\tau = 10$) and $\hat{S} = 1{,}340$ (SE $= 91$, $\tau = 6$) for S3a and 321, respectively. An advantage of this method, apart from its simplicity, is that estimates for f_0 can be produced under relatively weak conditions in which other methods might fail, and the estimates produced are remarkably consistent with those from other (successful) methods, even using relatively few contiguous frequencies, i.e., at relatively low τ (e.g., $\tau = 5$–10). However, modeling the logarithm

of the ratios, which is often required to ensure nonnegativity of implied f_0 estimates, is perhaps not ideal.

We are currently working on generalizing this approach from $(j+1)f_{j+1}/f_j$ being linear to f_{j+1}/f_j being a ratio of polynomials. Under this extended approach, the true data structure appears to be well captured, weighting can be employed to account for heteroscedasticity, and estimates for f_0 have been found to be positive without transformation. Model selection techniques suggest that the parsimonious model $f_{j+1}/f_j \approx (b_0 + b_1 j)/(a + j)$ (where b_0, b_1, and a are constants) may capture the data structure without the need to introduce higher-order polynomial terms. Interestingly, a mild generalization to the Katz structure (where $a = 1$) performs well even compared with models having higher-order terms. Tripathi & Gurland (1977) investigated this particular recursive distribution structure (for general a) and characterized the corresponding distribution by the property that its probability-generating function is a ratio of certain hypergeometric functions. This model appears to satisfy the objectives of simplicity and parsimony while producing sensible estimates and (potentially) fitting the data well. Another advantage of these regression-type methods is that they incorporate all frequencies up to the first noncontiguous counts ($f_j = 0$), whereas other estimation techniques often truncate the data artificially at an arbitrary cutoff τ, as discussed above. Disadvantages include difficulties associated with goodness-of-fit testing and model selection for nonlinear models, along with selection of the appropriate weighting scheme to account for heteroscedasticity. Models fitted with weighting $1/x$ give the total diversity estimate as 1,800 for S3a ($\tau = 31$) and 1,557 for 321 ($\tau = 33$); **Figure 3** shows the fitted curves. We do not yet have an SE calculation for this method.

4.2. More General Models

In the effort to better fit real, naturally occurring frequency count data sets, researchers have ranged ever further afield in their proposals for species abundance (mixing) distributions F, e.g., the generalized inverse Gaussian and finite mixtures of gammas (Section 3.3). Given the basic structure of the mixed Poisson model, however, the observed counts will always be marginally (zero-truncated) i.i.d. mixed Poisson random variables. Rather than looking for more complex mixing distributions, we can question both the independence and marginal distributional assumptions. On the latter point, Puri & Goldie (1979) characterized the extent of the mixed Poisson class (essentially, every derivative of the probability-generating function must be finite, even on the negative half-line), and they gave a counterexample in which the probabilities $p(j)$ depend on Pochhammer polynomials. Such distributions are worth investigating for species counts, in particular because their tail behavior differs from that of mixed Poissons, but to date, no published research in this vein exists.

However, investigation of alternative marginal distributions does not address the independence assumption. One way to view the standard model is as follows: There are S independent Poisson processes $X_1(t), X_2(t), \ldots, X_S(t)$ with rates $\lambda_1, \lambda_2, \ldots, \lambda_S$, independently contributing representatives (events) to the sample. The processes are stopped at some time t_0 (the sampling effort) and the collection counted to produce the frequency count data. From this perspective, we clearly may generalize either the marginal process type or the dependence structure, or both. In terms of altering the dependence structure, theory for multivariate (dependent) Poisson processes exists. Different constructions have been proposed (e.g., Karlis & Meligkotsidou 2007). Copulas have been used in this setting (Bäuerle & Grübel 2005): Write the joint distribution function Ψ of the S-dimensional random vector $(X_1(t), X_2(t), \ldots, X_S(t))$ as $\Psi(x_1, x_2, \ldots, x_S) = C(\Psi_1(x_1), \Psi_2(x_2), \ldots, \Psi_S(x_S))$, where the Ψ_i are the desired marginal Poisson distributions, and the copula, C, is an S-dimensional cumulative distribution function on the unit cube. To change both the marginal distributions and the

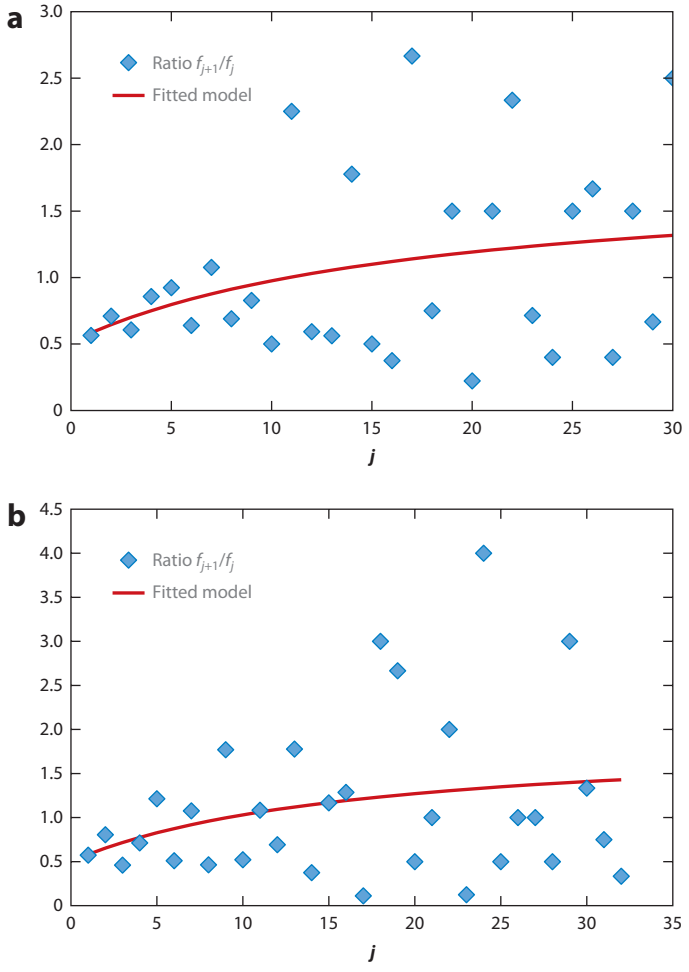

Figure 3

Ratio plots f_{j+1}/f_j as a function of j, with fitted ratio-of-linear-functions model, for (*a*) data set S3a and (*b*) data set 321.

dependence structure, the Poisson process may be generalized to a renewal process, and moment-based dependence may be introduced in the form of a covariance matrix Σ. Steinebach & Eastwood (1997) considered such multivariate renewal processes in an insurance problem. Another alternative for multivariate count data is the Sarmanov model from economics (Miravete 2009). For example, the Sarmanov structure for $S = 2$ is $\psi(x_1, x_2) = \psi_1(x_1)\psi_2(x_2)(1 + w\phi_1(x_1)\phi_2(x_2))$, where ψ is the probability mass function of the counts, w is a constant, and ϕ_1 and ϕ_2 are functions. We are not aware of any application of these models in the species problem to date.

5. ESTIMATING THE SAMPLE COVERAGE

If the population proportions of the species are $p_1, p_2, \ldots, p_S$, then the coverage is the total proportion of the population represented in the sample: $\sum_{i=1}^{S} p_i 1(X_i > 0)$. In some applications, interest focuses mainly on estimating the coverage or some function thereof. (The coverage is a

random variable, so technically we speak of predicting rather than estimating it.) For example, the opposite quantity, 1 − coverage, is the probability of observing a new species or the probability that the next sample item will be a member of a hitherto unobserved species, which is of particular interest for planning new studies. By now, a sizeable literature has accumulated on this issue, again dating back to Turing, but we have elected here to regard it as tangential but not central to the species estimation problem; a review of coverage estimation would be a project of a size comparable to that of the present one. We confine ourselves to noting that (as in the diversity problem) the full spectrum of statistical approaches has been explored: parametric to nonparametric and frequentist to Bayesian. For recent entrées to the literature, we refer the reader to Favaro et al. (2012), Gao (2013), and Lladser et al. (2011).

6. β-DIVERSITY AND INFERENCE

Microbial ecologists have great interest in β-diversity, and certainly it is the next step in analysis after α-diversity. For example, in the Human Microbiome Project, mapping the behavior, in particular the diversity, of microbial communities to varying conditions such as medical interventions, e.g., administration of antibiotics, is of critical importance. Several methods have been developed for this purpose, which typically take into account both the identities of the species existing at different times or locations and their diversity (variously defined). These procedures include UniFrac (Lozupone et al. 2011, Holmes et al. 2012), methods of coinertia (Dray et al. 2003), and ordination methods such as correspondence analysis (e.g., Greenacre 2007); they are often combined with graphical techniques such as multidimensional scaling. However, such methods treat the observed sample as the population; they do not make inference about the unobserved part of the population and so fall outside our scope here, especially because they are well documented elsewhere.

We are aware of only one procedure for β-diversity that incorporates inference about the unseen number of species. This procedure comes from Chao and coauthors and interestingly has been applied to microbial diversity (soil ciliates) (see Pan et al. 2009 and references therein). Pan et al. provided a nonparametric lower bound for the number of species shared between two or more populations. For two populations, the bound is $\hat{S}_{12} = D_{12} + af_{1+}^2/2f_{2+} + bf_{+1}^2/2f_{+2} + abf_{11}^2/4f_{22}$, where D_{12} is the observed number of shared species, f_{jk} is the number of species observed j times in sample 1 and k times in sample 2, and a and b are constants. (Note the similarity in the structure of the above bound to that of the α-diversity lower bound, Chao1, mentioned above.) In addition, Hampton & Lladser (2012) addressed this problem from the coverage perspective, estimating "the probability of a draw from one distribution not being observed in k draws from another distribution" (p. 1). In a related development, Chao et al. (2006) provided estimators that account for unobserved species when making inference about population similarity indices such as the Jaccard index. But generally speaking, true statistical inference for β-diversity, defining diversity as species richness, appears to be in the early stages and represents fertile ground for development.

7. FUTURE DIRECTIONS

We see three main areas for research and development. First, as noted above, the standard mixed Poisson model, although useful and even powerful, is not adequate to capture all of the natural complexity of microbial diversity sampling. Generalizations of both the model and statistical approaches are discussed above. These areas are ripe for theoretical research. Second, true inference for β-diversity, taking into account the unobserved portion of the population, is only beginning to be explored. This exploration is crucial for applications because researchers, both in environmental microbiology and in human health studies, have a pressing need to relate microbial

diversity to metadata or predictor variables, such as biogeochemical conditions or health-related interventions. Here, too, a great scope exists for theoretical research. Finally, at a practical level, intensive software development is necessary. We do not currently have an available platform on which the applied researcher can run multiple competing analyses of all types described above, along with corresponding diagnostic and graphical assessments. (CatchAll is a step in this direction, incorporating a particular suite of parametric mixed Poisson models and the coverage-based nonparametric estimators, but as we have discussed, many other powerful but unimplemented methods exist.) Furthermore, future software must go beyond mere graphics or even scientific visualization to visual analytics, which "closely integrates computational analysis and visualization and human-computer interaction" (Foster et al. 2012, p. 425). Visual analytics has been used in microbial ecology but not in conjunction with efforts to account for unobserved species. The combination of theoretical development with powerful applied tools will take the field of microbial diversity analysis to the next level.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We thank the Editors for proposing this review and an anonymous referee for careful reading and constructive comments.

LITERATURE CITED

Allen HK, Bunge J, Foster JA, Bayles DO, Stanton TB. 2013. Estimation of viral richness from shotgun metagenomes using a frequency count approach. *Microbiome* 1:5

Amann R, Fuchs BM, Behrens S. 2001. The identification of microorganisms by fluorescence in situ hybridisation. *Curr. Opin. Biotechnol.* 12:231–36

Barger K, Bunge J. 2011. Objective Bayesian estimation for the number of species. *J. Bayesian Anal.* 5:765–86

Bäuerle N, Grübel R. 2005. Multivariate counting processes: copulas and beyond. *ASTIN Bull.* 35:379–408

Berger JO, Bernardo JM, Sun D. 2012. Objective priors for discrete parameter spaces. *J. Am. Stat. Assoc.* 107:636–48

Bhat S, Sproat R. 2009. Knowing the unseen: estimating vocabulary size over unseen samples. In *Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and 4th International Joint Conference on Natural Language Processing of the AFNLP: Proceedings of the Conference*, Vol. 1, pp. 109–117. Stroudsburg, PA: World Sci. Publ.

Böhning D, Kuhnert R. 2006. Equivalence of truncated count mixture distributions and mixtures of truncated count distributions. *Biometrics* 62:1207–15

Böhning D, Kuhnert R. 2009. CAMCR: Computer-Assisted Mixture model analysis for Capture-Recapture count data. *AStA Adv. Stat. Anal.* 93:61–71

Böhning D, Schön D. 2005. Nonparametric maximum likelihood estimation of population size based on the counting distribution. *J. R. Stat. Soc. C* 54:721–37

Bunge J. 2013. A survey of software for fitting capture-recapture models. *WIREs Comput. Stat.* 5:114–20

Bunge J, Barger K. 2008. Parametric models for estimating the number of classes. *Biom. J.* 50:971–82

Bunge J, Böhning D, Allen H, Foster JA. 2012a. Estimating population diversity with unreliable low frequency counts. In *Biocomputing 2012: Proceedings of the Pacific Symposium*, pp. 203–12. Hackensack, NJ: World Sci. Publ.

Bunge J, Fitzpatrick M. 1993. Estimating the number of species—a review. *J. Am. Stat. Assoc.* 88:364–73

Bunge J, Woodard L, Böhning D, Foster JA, Connolly S, Allen HK. 2012b. Estimating population diversity with CatchAll. *Bioinformatics* 28:1045–47

Chao A. 1987. Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* 43:783–91

Chao A. 2005. Species richness estimation. In *Encyclopedia of Statistical Sciences*, ed. S Kotz, N Balakrishnan, CB Read, B Vidakovic, pp. 7907–16. New York: Wiley. 2nd ed.

Chao A, Chazdon RL, Colwell RK, Shen T-J. 2006. Abundance-based similarity indices and their estimation when there are unseen species in samples. *Biometrics* 62:361–71

Chao A, Shen T-J. 2003–2005. Program SPADE (Species Prediction and Diversity Estimation). *Stat. Softw.* Program and user's guide at **http://chao.stat.nthu.edu.tw**

Dray S, Chessel D, Thioulouse J. 2003. Co-inertia analysis and the linking of ecological data tables. *Ecology* 84:3078–89

Favaro S, Lijoi A, Pruenster I. 2012. A new estimator of the discovery probability. *Biometrics* 68:1188–96

Fisher RA, Corbet S, Williams CB. 1943. The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Anim. Ecol.* 12:42–58

Foster JA, Bunge J, Gilbert JA, Moore JH. 2012. Measuring the microbiome: perspectives on advances in DNA-based techniques for exploring microbial life. *Brief. Bioinforma.* 13:420–29

Gao F. 2013. Moderate deviations for a nonparametric estimator of sample coverage. *Ann. Stat.* 41:641–69

Gilbert JA, O'Dor R, King N, Vogel TM. 2011. The importance of metagenomic surveys to microbial ecology: or why Darwin would have been a metagenomic scientist. *Microb. Inform. Exp.* 1:5

Good IJ. 1953. The population frequencies of species and the estimation of population parameters. *Biometrika* 40:237–64

Greenacre M. 2007. *Correspondence Analysis in Practice*. London: Chapman & Hall/CRC. 2nd ed.

Hampton J, Lladser ME. 2012. Estimation of distribution overlap of urn models. *PLoS ONE* 7:e42368

Holmes I, Harris K, Quince C. 2012. Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PLoS ONE* 7:e31026

Johnson NL, Kemp AW, Kotz S. 2005. *Univariate Discrete Distributions*. Hoboken, NJ: Wiley

Karlis D, Meligkotsidou L. 2007. Finite mixtures of multivariate Poisson distributions with application. *J. Stat. Plan. Inference* 137:1942–60

Koenker R, Mizera I. 2013. Convex optimization in R. *J. Stat. Softw.* In press

Lewis K. 2009. Persisters, biofilms, and the problem of cultivability. In *Uncultivated Microorganisms*, ed. S Epstein, pp. 181–94. Berlin: Springer-Verlag

Lladser ME, Gouet R, Reeder J. 2011. Extrapolation of urn models via Poissonization: accurate measurements of the microbial unknown. *PLoS ONE* 6:e21105

Logares R, Haverkamp TH, Kumar S, Lanzén A, Nederbragt AJ, et al. 2012. Environmental microbiology through the lens of high-throughput DNA sequencing: synopsis of current platforms and bioinformatics approaches. *J. Microbiol. Methods* 91:106–13

Lozupone C, Lladser ME, Knights D, Stombaugh J, Knight R. 2011. UniFrac: an effective distance metric for microbial community comparison. *ISME J.* 5:169–72

Mao CX, Colwell RK. 2005. Estimation of species richness: mixture models, the role of rare species, and inferential challenges. *Ecology* 86:1143–53

Mao CX, Lindsay BG. 2007. Estimating the number of classes. *Ann. Stat.* 35:917–30

Miravete E. 2009. *Multivariate Sarmanov count data models*. CEPR Discuss. Pap. 7463, Cent. Econ. Policy Res., London

Ohannessian MI, Dahleh MA. 2012. *Large alphabets: finite, infinite, and scaling models*. Presented at the 46th Annu. Conf. Inf. Sci. Syst. (CISS), Inst. Electr. Electron. Eng., Princeton, NJ

Pan HY, Chao A, Foissner W. 2009. A nonparametric lower bound for the number of species shared by multiple communities. *J. Agric. Biol. Environ. Stat.* 14:452–68

Puri PS, Goldie CM. 1979. Poisson mixtures and quasi-infinite divisibility of distributions. *J. Appl. Probab.* 16:138–53

Quince C, Curtis TP, Sloan WT. 2008. The rational exploration of microbial diversity. *ISME J.* 2:997–1006

Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ. 2011. Removing noise from pyrosequenced amplicons. *BMC Bioinforma.* 12:38–55

Rocchetti I, Bunge J, Böhning D. 2011. Population size estimation based upon ratios of recapture probabilities. *Ann. Appl. Stat.* 5:1512–33

Sears CL. 2005. A dynamic partnership: celebrating our gut flora. *Anaerobe* 11:247–51

Staley JT, Konopka A. 1985. Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annu. Rev. Microbiol.* 39:321–46

Steinebach J, Eastwood VR. 1997. Detecting changes in a multivariate renewal process. *Metrika* 46:1–19

Tardella L. 2002. A new Bayesian method for nonparametric capture-recapture models in presence of heterogeneity. *Biometrika* 89:807–17

Tripathi RC, Gurland J. 1977. A general family of discrete distributions with hypergeometric probabilities. *J. R. Stat. Soc. B* 39:349–56

Tuomisto H. 2011. Commentary: Do we have a consistent terminology for species diversity? Yes, if we choose to use it. *Oecologia* 167:903–11

Valero J, Pérez-Casany M, Ginebra J. 2010. On zero-truncating and mixing Poisson distributions. *Adv. Appl. Probab.* 42:1013–27

Valiant G, Valiant P. 2011. Estimating the unseen: an n/log(n)-sample estimator for entropy and support size, shown optimal via new CLTs. In *Proceedings of the 43rd Annual ACM Symposium on Theory of Computing*, ed. L Fortnow, SP Vadhan, pp. 685–94. New York: ACM

Wang J-P. 2010. Estimating species richness by a Poisson-compound gamma model. *Biometrika* 97:727–40

Whitman WB, Coleman DC, Wiebe WJ. 1998. Prokaryotes: the unseen majority. *Proc. Natl. Acad. Sci. USA* 95:6578–83

Williamson M, Gaston KJ. 2005. The lognormal distribution is not an appropriate null hypothesis for the species-abundance distribution. *J. Anim. Ecol.* 74:409–22

Zhang Z. 2012. Entropy estimation in Turing's perspective. *Neural Comput.* 24:1368–89

Zhang Z, Zhou J. 2010. Re-parameterization of multinomial distributions and diversity indices. *J. Stat. Plan. Inference* 140:1731–38