STATISTICAL ANALYSIS OF PYROSEQUENCE DATA

by

KAREN KEATING

B.S., Arkansas State University, 1977
M.A., University of Maryland, 1981

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2012

Abstract

Since their commercial introduction in 2005, DNA sequencing technologies have become widely available and are now cost-effective tools for determining the genetic characteristics of organisms. While the biomedical applications of DNA sequencing are apparent, these technologies have been applied to many other research areas. One such area is community ecology, in which DNA sequence data are used to identify the presence and abundance of microscopic organisms that inhabit an environment. This is currently an active area of research, since it is generally believed that a change in the composition of microscopic species in a geographic area may signal a change in the overall health of the environment.

An overview of DNA pyrosequencing, as implemented by the Roche/Life Science 454 platform, is presented and aspects of the process that can introduce variability in data are identified. Four ecological data sets that were generated by the 454 platform are used for illustration. Characteristics of these data include high dimensionality, a large proportion of zeros (usually in excess of 90%), and nonzero values that are strongly right-skewed.

A nonparametric method to standardize these data is presented and effects of standardization on outliers and skewness are examined. Traditional statistical methods for analyzing macroscopic species abundance data are discussed, and the applicability of these methods to microscopic species data is examined. One objective that receives focus is the classification of microscopic species as either rare or common species. This is an important distinction since there is much evidence to suggest that the biological and environmental mechanisms that govern common species are distinctly different than the mechanisms that govern rare species. This indicates that the abundance patterns for common and rare species may follow different probability models, and the suitability of the Pareto distribution for rare species is examined. Techniques for classifying macroscopic species are shown to be ill-suited for microscopic species, and an alternative technique is presented. Recognizing that the structure of the data is similar to that of financial applications (such as insurance claims and the distribution of wealth), the Gini index and other statistics based on the Lorenz curve are explored as potential test statistics for distinguishing rare versus common species.

STATISTICAL ANALYSIS OF PYROSEQUENCE DATA

by

KAREN KEATING

B.S., Mathematics, Arkansas State University, 1977
M.A., Applied Mathematics, University of Maryland, 1981

A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2012

Approved by:

Major Professor
Gary L. Gadbury

Abstract

Since their commercial introduction in 2005, DNA sequencing technologies have become widely available and are now cost-effective tools for determining the genetic characteristics of organisms. While the biomedical applications of DNA sequencing are apparent, these technologies have been applied to many other research areas. One such area is community ecology, in which DNA sequence data are used to identify the presence and abundance of microscopic organisms that inhabit an environment. This is currently an active area of research, since it is generally believed that a change in the composition of microscopic species in a geographic area may signal a change in the overall health of the environment.

An overview of DNA pyrosequencing, as implemented by the Roche/Life Science 454 platform, is presented and aspects of the process that can introduce variability in data are identified. Four ecological data sets that were generated by the 454 platform are used for illustration. Characteristics of these data include high dimensionality, a large proportion of zeros (usually in excess of 90%), and nonzero values that are strongly right-skewed.

A nonparametric method to standardize these data is presented and effects of standardization on outliers and skewness are examined. Traditional statistical methods for analyzing macroscopic species abundance data are discussed, and the applicability of these methods to microscopic species data are? examined. One objective that receives focus is the classification of microscopic species as either rare or common species. This is an important distinction since there is much evidence to suggest that the biological and environmental mechanisms that govern common species are distinctly different than the mechanisms that govern rare species. This indicates that the abundance patterns for common and rare species may follow different probability models, and the suitability of the Pareto distribution for rare species is examined. Techniques for classifying macroscopic species are shown to be ill-suited for microscopic species, and an alternative technique is presented. Recognizing that the structure of the data is similar to that of financial applications (such as insurance claims and the distribution of wealth), the Gini index and other statistics based on the Lorenz curve are explored as potential test statistics for distinguishing rare versus common species.

Table of Contents

List of Figures

List of Tables

Chapter 1. Introduction

1.1. Background of the Application

This research involves statistical methods development for the analysis of data from pyrosequencing technology. This technology, described in more detail later, allows for the study of species composition at the microscopic level and is one of the new high-throughput technologies to emerge in recent years. The research considered here is based on a premise that statistical methods development for pyrosequencing experiments will follow a similar path to those developed (and in progress) for other recent high-throughput technologies. One of these is the technology associated with high-throughput gene expression experiments using microarrays.

Microarrays measure the expression of thousands of genes simultaneously and are used to determine which genes are activated by certain stimuli (i.e., treatment conditions). The earliest experiments had a single sample (microarray) in each of two treatment groups and the log-ratio of gene expression across the two treatments was used to determine a "fold change" (*e.g.*, Lee *et al.*, 1999). An arbitrary cut-off was used to determine "significant results" or those genes with expressions that were altered by the treatment. Little was known about the meaning of the expression levels, how much of it was noise or technical artifacts resulting from the new technology, and whether there was bias in measurements. Issues related to the design of such experiments, corrections for multiple testing in high-dimensional data, and software tools to process such data had not been considered. In subsequent years, statisticians became involved and the number of publications with microarray and statistics as keywords grew substantially in the years after 2000. For example, the number of papers in Web of Science with '(Gene Expression OR Microarray) AND Statistics' as keywords was 65 for the five years 1995 – 1999, 321 for 2000 – 2004, and 938 for 2005 – 2009. The early statistics-related papers introduced the microarray technology from a statistical perspective and identified statistical challenges related to the new technology (*e.g.*, Zhang, 1998). Other papers concerned correction for background noise, and others normalization of data to remove systematic sources of bias and variation introduced by the technologies (*cf.*, Irizarry *et al.*, 2003a). Many papers concerned experimental designs for microarray experiments (*e.g.* Simon and Dobbin, 2003), others sample size requirements (*e.g.*, Gadbury *et al.*, 2004), and many more the multiple testing problem where thousands of hypotheses are being tested simultaneously (*e.g.,* Storey, 2002). Research is

ongoing and new microarray platforms introduce new issues to be considered in statistical analyses.

The new area of pyrosequencing seems poised for methods development paralleling what happened with the emergence of microarrays though, as we will see, the statistical challenges are different, as are the applications that motivate pyrosequencing experiments. While true that pyrosequencing experiments have been designed, data collected and analyzed, and papers published, literature reviews suggest less evidence that statisticians have become engaged. This research thus represents a foray into this emerging field, and is intended to offer a beginning and to motivate follow-on research activity in this exciting and important area.

To appreciate the motivating application and the nature of pyrosequencing experiments and data, this introduction concludes with a background of community ecology studies, pyrosequencing data, and then a subsection with more detail on pyrosequencing technology. Then, in Chapter 2, current methods of analyzing pyrosequencing data are reviewed, and some additional description of data provided. Four ecological pyrosequence data sets are introduced in Chapter 3, and exploratory data analysis reveals the common structure of these types of data. In Chapter 4, two new methodologies are presented, which comprise the central results of this research. These are a nonparametric procedure to standardize the data and a new method for classifying common and rare species. Chapter 5 concludes with a summary of the results of this research and outlines potential areas of future research.

1.1.1. Community Ecology

In the field of ecology, a community is a group of interacting species that inhabit a particular location at a particular time. Community ecologists study the collection of species in a community and investigate factors that affect the collection. Comparing two or more communities could involve multiple locations, or could be the same location at different times. Factors that affect a community's structure may be related to the natural environment (*e.g.* prairie, forest) or they may be related to the species themselves (*e.g.* predators, prey). In some cases, the factor of interest is time. This would occur, for example, when the ecologist is studying the recovery of a habitat after a natural disaster such as a forest fire or volcanic eruption.

To describe communities, ecologists typically use measures such as species richness, species evenness, and diversity. Species richness is simply the number of distinct species in a community. Species evenness measures the equity of the abundances across the species. Species evenness is highest when all species in a community have the same abundance and approaches zero as a single species becomes more dominant. Diversity is more complex, and incorporates both species richness and evenness. There are different methods for calculating species richness, evenness and diversity. These are described in Chapter 2.

Customarily, community ecologists study a narrow group of species and/or environments. For example, Fischer *et al.* (2011) studied bird communities in agricultural landscapes and Magurran and Henderson (2003) examined a fish community across 21 years at a particular location in the United Kingdom. These studies involve macroscopic organisms, ones that can be visually observed and assigned to a particular taxonomic category (a species). It is now recognized that species can be identified by their genetic material. This allows community ecologists to examine microscopic species, such as fungi and bacteria, in order to obtain a more complete picture of community dynamics.

1.1.2. Pyrosequencing

Pyrosequencing is one method for interrogating the genetic material of an organism. It can be used for two primary purposes: to gain a complete picture of the genetic material of an organism (its genome) and to identify the subtle differences in genomes that make each individual unique. For example, the human genome project was completed in 2003 and, among other things, identified the precise chemical composition of human DNA (deoxyribonucleic acid) comprising the genome. Subsequent interrogation of human DNA has identified variants responsible for specific human diseases. According to the National Human Genome Research Institute in the National Institutes of Health (www.genome.gov/25521731), the genomes of many less complex organisms have also been mapped, including E. coli, baker's yeast, the roundworm, and the fruit fly. Comparison of these genomes to the human genome provide insights into evolutionary history and aid in identifying the portions of the genome responsible for requisite biological functions of the organisms.

Pyrosequencing can also be used to identify the taxonomic classification (*e.g.* species) of an organism. Prior to DNA analysis, taxonomic classification relied on morphological features

such as shape, color, size and behavior. Distinguishing between closely related species was problematic, since they could share the same morphology. Examining the genomes of organisms provides a less ambiguous method of classification, and also allows the classification of microscopic organisms for which morphology is less precise.

1.1.3. Nomenclature

This research is motivated by ecological data sets generated via pyrosequencing. In ecological applications, each analyte (for example, a soil core or a leaf) is called a sample. Each analyte represents an ecological community, that is, a collection of interacting species at a particular location and time, so the analytes are also called sites or plots. For the remainder of this document, the terms sample, site, plot and analyte are all synonymous. Each sample is characterized by a vector of counts indicating the DNA sequences that were identified in the analyte. Some researchers use the term Operational Taxonomic Unit (OTU), while others use the term cluster. These both refer to a (nearly) unique DNA sequence measured from an analyte, and it is assumed that these are a surrogate for species. Thus the terms OTU, cluster and species are treated as synonyms. The counts in each data set are derived by binning the observed DNA sequences into similar sequence patterns. Thus the counts are sometimes referred to as the number of sequences or the number of reads. Collectively, the counts are called abundances. Relative abundances, also called relative frequencies, are the proportional abundances of each OTU at each site, so that the total relative abundance for each site is equal to one.

In initial investigations of ecological OTU data, we are interested in certain summaries. The total count for a site (across all OTUs) is called a site total. The total count for an OTU (across all sites) is called an OTU total. A count for a particular OTU at a particular site is called an individual OTU count.

1.2. Pyrosequence Data

1.2.1. Overview of Pyrosequencing

Pyrosequencing was developed in the 1990's by Mostafa Ronaghi, Pål Nyrén, Mathias Uhlén (1998) in collaboration with several colleagues and graduate students. They sold the first automated system in 1999 (Nyrén, 2007), and received a U.S. patent in 2001. There are now several commercial platforms for pyrosequencing, commonly referred to as 'next generation' or

'massively parallel' sequencing. These platforms include Roche 454 (http://www.454.com), Illumina (http://www.illumina.com), Helicos (http://www.helicosbio.com), and SOLiD (http://www.appliedbiosystems.com). These systems, among others, all rely on a complex combination of chemistry and computing capabilities. While these platforms may differ in the details of their implementation, they all utilize a procedure known as pyrosequencing, or sequencing by synthesis. In the description that follows, we focus on the Roche 454 platform since it is data from this sequencer that we will be examining.

The pyrosequencing reaction relies on the chemical bonds that occur within DNA's double helix. The structure is a double-stranded chain consisting of paired nucleotides, and the ordering of the nucleotides in the chain define the genetic information. DNA naturally occurs as a double strand, and the nucleotides on one strand are complementary to the nucleotides on the second strand. The complementary nucleotides are known as base pairs; for DNA these are adenine-thymine (A-T) and guanine-cytosine (G-C). For example, if one DNA strand contains the nucleotides ATTCG then the complementary strand is TAAGC. The chemical bonding that occurs between the two strands ensures that the nucleotides will be complementary. Unless the DNA has mutated or has been damaged (for example, by radiation), every nucleotide on one strand is paired with its complement on the second strand. Pyrosequencing relies on the natural formation of these complementary base pairs.

To prepare an analyte for pyrosequencing, its DNA must be extracted. Particular sections of the DNA that are pertinent to the current research objective are isolated and extracted. Each section is molecularly tagged to identify the analyte, then attached to a small bead and duplicated via polymerase chain reaction (PCR) amplification. After PCR, each bead holds millions of single-stranded copies of the original fragment. These beads are placed into wells on a picotiter plate, one bead per well, and are ready for pyrosequencing. This process is illustrated in Figure 1.1.

When the beads are loaded onto the plate, the process of pyrosequencing can begin. This process is automated and occurs within the sequencer. The plate is flooded with a solution that contains one of the four types of nucleotide, plus enzymes and other reagents to control the chemical reactions. The DNA strands attached to the beads will incorporate this nucleotide only if they are complementary base pairs. If this happens, a small burst of light is emitted and the luminescence is captured by the pyrosequencer. The excess solution is removed, and the plate is

flooded again, but this time with a different nucleotide. Again, any luminescence is recorded by the pyrosequencer, and the process is repeated a fixed number of times. Each flooding process is called a flow, and a series of four flows (using each of the four nucleotides) is called a flow cycle. Since the strands on a bead will incorporate only a complementary nucleotide, knowing which nucleotide is incorporated tells us which nucleotide is on the original single strand. Thus the sequence of nucleotides on the original DNA fragment can be identified.

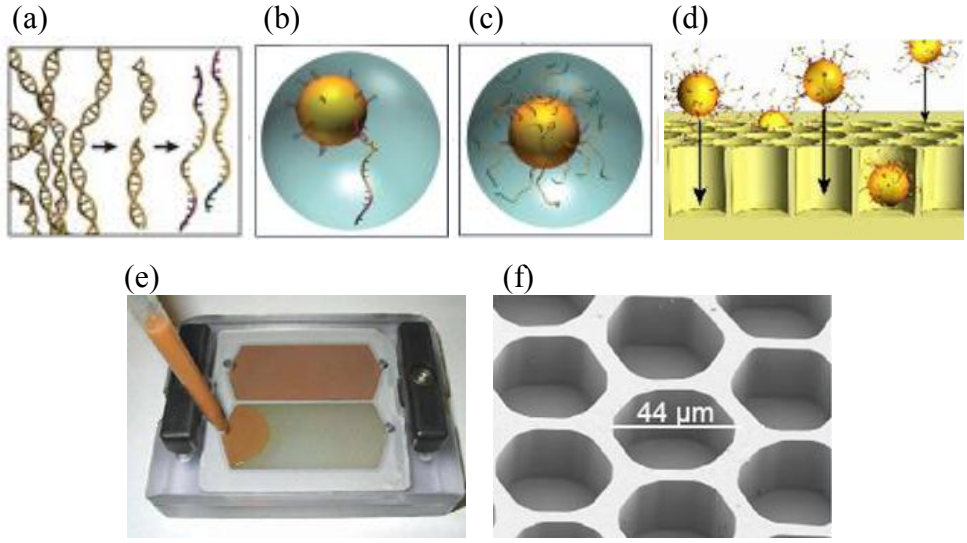


Figure 1.1: Preparing an Analyte

Target DNA is extracted (a), and attached to bead (b). After PCR amplification, bead contains millions of duplicates (c) and is placed into a well on a picotiter plate (d). Picotiter plate being loaded (e) and magnified (f). The plate contains 1.6 million wells and each well is approximately one-third the width of a human hair. Image source: http://www.454.com.

When a strand incorporates a nucleotide, the amount of light emitted is proportional to the number of nucleotides incorporated. The pyrosequencer records both the order in which nucleotides are applied to the plate and the amount of light emitted at each well. The result for each well is called a 'read' and can be displayed in the form of a pyrogram (also called a flowgram), as shown in Figure 1.2. There is one pyrogram (one read) for each well, and 1.6 million wells on a plate.

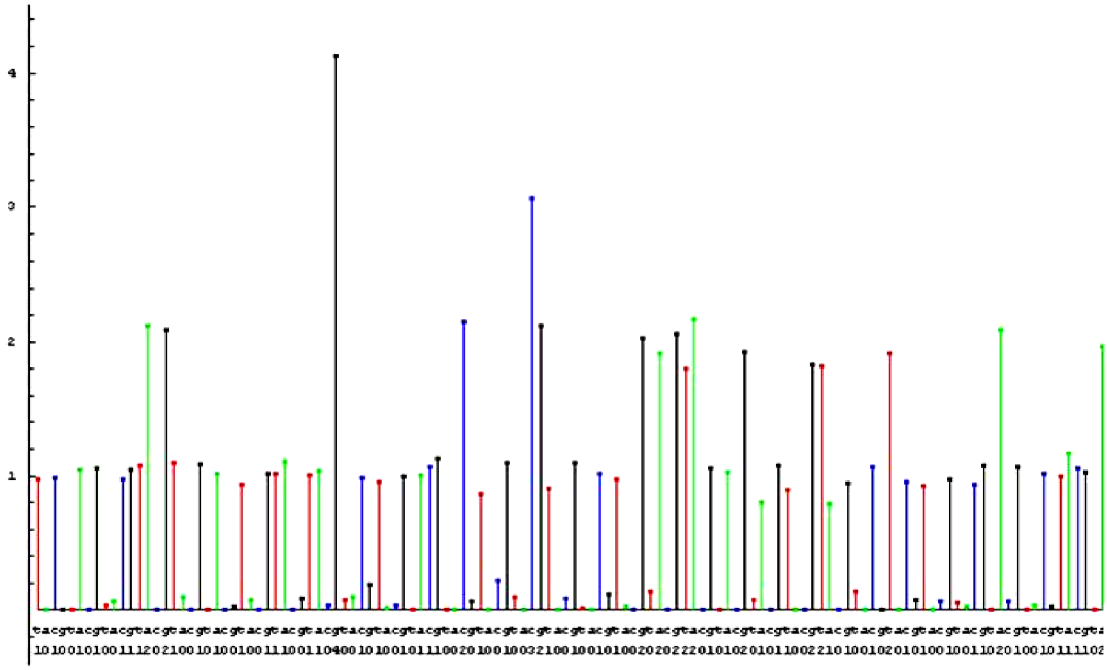


Figure 1.2: Example Pyrogram

Letters along the x-axis indicate which nucleotide is washed over the plate and the height indicates the intensity of illumination, which is proportional to the number of nucleotides incorporated. Reading from the left, the first few incorporated nucleotides are TCAGCGTAAGG, so the DNA strand in this well begins with AGTCGCATTCC.
Image source: http://www.pmgf.osu.edu/services_mps.html.

Thus the data from the pyrosequencer consist of an enormous collection of fragmented DNA sequences. Similar fragments are clustered together to form Operational Taxonomic Units (OTUs). There are numerous mathematical approaches and software implementations for clustering these fragments, and research in this area is ongoing. Two commonly used programs are CAP3 (Huang and Madan, 1999) and Pyrotagger (Kunin and Hugenholtz, 2010a). At the conclusion of the clustering process, the pyrosequence data consist of a collection of OTUs along with the frequency of each OTU in each analyte. The recorded data for each analyte is a vector of counts and the elements in the vector correspond to OTUs. At this stage, the pyrosequence data can be envisioned as a matrix in which the rows represent analytes and the columns represent OTUs. The entries in the matrix consist of counts, *i.e.* the number of times each OTU was identified in each analyte.

1.2.2. Sources of Variability

Data generated via pyrosequencing have many unique characteristics. Typically, the number of analytes is in the hundreds (or less) and the number of OTUs is in the thousands. The proportion of zeros is large, usually in excess of 90%, and the nonzero counts are highly variable. There is variability in the total count for each analyte, and in the total count for each OTU. In one typical data set (the soil data described in Chapter 3), the OTU total counts range from 1 to 14749, with mean 64 and standard deviation 482. The site total counts range from 7 to 2929, with mean 648 and standard deviation 364. To adequately analyze these data, it is necessary to identify and model the sources of this variability so that remaining variability can be attributed to experimental and/or environmental conditions. The major sources of variability are described below. They are identified in the order they occur in the workflow, since variability at any stage of the process affects all downstream analysis.

1.2.2.1. DNA Extraction and PCR Amplification

Short strands (15 to 30 nucleotides) called primers are used to isolate the target DNA for PCR amplification. There are two general kinds of primers: one to identify the beginning of the target DNA strand and one to identify the end. The nucleotides in the primers are complementary to the beginning and end of the target DNA, so they are genome-specific and target-specific. It is the sequence in the primers that determine where they bind to the sample DNA and therefore define the section of DNA to be amplified. Selection of an appropriate primer requires knowledge of the genome, and new primers are synthesized as more is learned about a genome. Small variations in the sample DNA surrounding the target area can affect the ability of the primer to bind to the site, which inhibits PCR amplification. Bellemain (2010) examined seven primers for fungal DNA (targeting the ITS region of the genome) and concluded "the selected ITS primers showed large variation in the ability to amplify fungal sequences" (page 4). When allowing one mismatch in the primer sequence, one primer amplified only 65% of the target DNA, while another amplified 91%. This disparity extends to taxonomic groups as well. One primer combination (ITS3-ITS4) amplified over 98% of the Ascomycetes sequences, but less than 74% of the Basidiomycetes. These two taxonomic groups comprise 79% of all species of fungi, so any discrepancy in their amplification rate could have a critical impact on the resulting OTU counts.

1.2.2.2. Base Calling

Pyrosequencing is accomplished by iteratively flowing each of the four nucleotides (A, C, G and T) over a picotiter plate containing wells of DNA to be sequenced. Base calling refers to the determination of how many, if any, nucleotides are incorporated in each well on each flow. Each incorporation emits a small amount of light so that, in theory, the amount of light produced at each well is proportional to the number of nucleotides incorporated. Thus a continuous value (the flow value, *i.e.* light intensity) must be translated into an integer (number of nucleotides). This action is known as base calling. Typically, the continuous values are simply rounded to whole numbers (Quince *et al*., 2009). Base call accuracy is directly dependent on the accuracy of the measured light intensity. One reason for performing PCR amplification is to increase this intensity, thus variation in amplification can affect base call accuracy. The most problematic issue with base calling stems from homopolymers, chains of consecutive identical nucleotides on the target DNA. For homopolymers less than 8 bases, the intensity of the light signal is linearly related to the number of nucleotides, but the signal degrades for longer homopolymers (Margulies *et al*., 2005), resulting in base calls that are too short. This can cause mismatches in what are supposed to be two identical fragments, which creates difficulties in fragment clustering and subsequent OTU identification.

A detailed examination (Gilles *et al*., 2011) of the Roche 454 GS FLX technology uncovered patterns in the base call error rates that can be attributed to specific sources. An 'edge effect' occurs because the light-sensing camera is located at the center of the plate, so that the light measurement from a well along an edge of the plate is not as accurate as one in the center. The 'direction effect' is a result of the direction in which the nucleotide solutions are flowed over the plate, so that wells which receive the solution first are more likely to have stronger light signals. Incomplete cleansing of the plate between flows can cause a 'carry forward' effect in which nucleotides from a previous flow remain in a well and are incorporated during the next flow. There can also be 'incomplete extensions' in which some of the strands on a bead fail to incorporate during the appropriate base flow. And finally, the position of the base along the target strand can also affect the accuracy of the base call, specifically, accuracy decreases as the length of the target DNA increases.

The precise characteristics and severity of these errors can be platform-specific, so there is no 'one size fits all' solution. Each sequencing platform contains built-in software for base

calling, and generates sequence data both in flow format and base call format. For each well, flow format contains the numeric values related to the light intensity at each flow, while base call format consists of a sequence of letters (*e.g.* TAACC) that represent the resulting base calls. There are numerous commercial and open-source software packages for base calling that can be used to bypass the sequencer's built-in program. These packages utilize the flow data generated by the sequencer, and incorporate more sophisticated techniques for assigning the bases. These packages include Phred (Beguelin and Nutt, 1994), naiveBayesCall (Kao and Song, 2011), BayesCall (Kao *et al.*, 2009), PyroBayes (Quinlan *et al.*, 2008), and Pyronoise (Quince *et al.*, 2009). In addition to assigning the bases, base calling programs also provide a 'quality score' for each base call, which can be used as a measure of reliability.

1.2.2.3. Contamination

As with any complex laboratory procedure, there is potential for mistakes, although strict adherence to protocols minimizes the risk. Ideally, the target DNA fragments should be long enough so that the end of the fragment is not reached during sequencing. If a fragment is too short, then part of the primer and other molecular tags attached to the fragment will become part of the sequenced read for the fragment. Thus the calculated sequence for the fragment is contaminated with the primer/tag sequence, which creates difficulties when trying to cluster the fragments. Because the wells are in such close proximity, there is also the possibility of crosstalk between the wells, that is, a reaction in one well can have an effect on nearby wells. This can be manifested as background noise in the flow values. These and other sources of potential contamination can be corrected by quality-trimming and read-filtering algorithms. (Balzer *et al.*, 2011) Quality trimming involves removing a portion (usually the end) of a read to eliminate a primer sequence and/or disregard the less accurate base calls. Read filtering refers to removal of an entire read for reasons such as too many 'dubious' flow values in the interval [0.5, 0.7], which indicates a low quality read. (Margulies *et al.*, 2005)

1.2.2.4. Clustering the Fragments

Prior to extracting the target DNA strands from an analyte, there are many copies of each DNA target because each cell in the analyte contains DNA. The primer extracts specific sections of the genome and each section is amplified and sequenced. Clustering algorithms

identify 'similar' sequenced sections and collect these into a set of Operational Taxonomic Units (OTUs).

The clustered sequences are rarely identical. The user provides thresholds that define how similar the sequences need to be in order to be clustered. There are several types of thresholds, but two of the most common are minimum overlap and percent similarity. The minimum overlap specifies the minimum number of contiguous nucleotides that two fragments have in common and the percent similarity defines a minimum percent of nucleotides that must match. For each of these thresholds, a higher value results in more stringent clustering requirements, so that fewer fragments are clustered. This results in a larger number of OTUs and smaller counts for these OTUs. Lower threshold values produce fewer OTUs, but the counts for the OTUs are higher. Regardless of the threshold values, it is customary to have several sequences that are not clustered at all. These result in individual OTU designations, each with a frequency of one. These are called singletons, and may represent an extremely rare species, but more likely these are the result of sequencing errors (Unterseher, *et al*., 2011).

Chapter 2. Current Methods

Statistical methods for summarizing and analyzing ecological OTU abundance data have generally been adapted from the methods used for analyzing macroscopic species. In general, species data are collected at several sites, and the research objective is to determine whether the species assemblage is different in different sites. This is a very broad question, and is often interpreted to mean: Does the *diversity* change across the sites? There are many types of diversity, including alpha (within-site) diversity and beta (between-site) diversity. In either case, diversity is a single numeric value. Larger values of α-diversity indicate a healthier habitat, whereas larger values of β-diversity indicate that the two sites are dissimilar.

Both α- and β-diversity are generally measured as a combination of species richness and species evenness. Species richness is simply the number of distinct species in the habitat and species evenness measures the equality with which the species are distributed in the habitat. While these are easy to explain, they are very difficult to measure. For both macroscopic species data and microscopic OTU data, true species richness must be estimated from observed species richness. Some of the methodologies are discussed in Section 2.2.

Statistical tests to determine if there are differences between sites can be formulated in terms of any one of several univariate measures of diversity, or can be based on the entire vector of abundances measured at each site. Furthermore, the vectors and the corresponding diversity measures can be based on raw abundances (actual counts), or proportions (percents of the site total). Wharton and Hui (2011) reviewed articles published in *Ecology* during 2008-2009 and discovered over one-third analyzed proportions rather than raw counts. Of these, the "most common method of analysis was to utilize the arcsine square root transform ... followed by a linear model." (page 3) They argue that the arcsine transform should no longer be used, and that it should be replaced by logistic regression or generalized linear mixed models.

This chapter summarizes methods that have been used for OTU data that have mostly appeared in ecological journals. It highlights steps taken to adapt relevant methods for analysis of macroscopic species to microscopic species. Some discussion is given on how the methods either do, or do not, seem to work with the distributional complexities of data from pyrosequencing experiments.

2.1. Characteristics of OTU Data

OTU data derived from pyrosequencing share many characteristics with macroscopic species abundance data, but there are several differences that make OTU data unique. For macroscopic organisms, such as birds or frogs, an individual organism is visually observed and classified into a taxonomic category (*i.e.*, a species). The resulting species abundance data consists of the number of observed individuals in each species. With microscopic OTU data, the quantity being measured is the number of PCR-amplified DNA fragments, which may or may not directly translate into the number of individuals. In fact, with microscopic organisms (such as bacteria), the biological concept of an 'individual' is not straightforward. Nevertheless, the questions we attempt to answer with these data remain the same. In general, we wish to determine if there are any environmental and/or experimental conditions that affect the occurrence or abundance of the organisms. To explore this question, we need to understand the basic characteristics of OTU data and how these data compare to macroscopic species data.

For both macroscopic species and OTUs, an abundance data set can be envisioned as a matrix in which the rows represent sites (*aka* analytes or samples) and the columns represent species (or OTUs). For macroscopic species, the size of the data set is directly dependent on the sampling effort and the scope of the investigation, but the number of species is typically in the hundreds or less. For microscopic data, the matrix contains a thousand or more OTUs, and the number of sites is in the hundreds or smaller. This presents unique challenges for analyzing OTU data, since there is insufficient information to model each OTU separately. In addition, OTU data sets contain a large proportion of zeros, usually in excess of 90%, and the distribution of nonzero values is strongly right-skewed.

There are biological models, for example niche apportionment models, to explain the abundances of macroscopic species. Many of the statistical distributions used to model macroscopic species abundance data are based on these biological models. It is unknown whether the biological and stochastic processes that govern macroscopic species abundances can be equated to the processes that generate OTU abundances. This distinction has been recognized, but not resolved (Magurran, 2004).

Another unique characteristic of OTU data is that of singletons. Singletons occur when a measured DNA sequence is unlike any other detected sequence, so that it is not clustered with any other sequence. This produces an OTU with a count of one. The presence of singletons in a

data set can have a profound impact on the estimate of species richness, which in turn affects the estimate of diversity. While singletons can occur in macroscopic species data (when a particularly rare species is observed only once), singletons are much more prevalent in OTU data and thus warrant special attention.

2.2. Species Richness

Species richness is simply the number of unique species in a community. In macroscopic data, it is typical that the observed species richness is less than the true species richness. This can occur because a species is present, but not observed. The degree to which the distinct species are undercounted is related to the sampling effort and the rarity of the species. In OTU data, the presence of singletons leads to an over-estimate of species richness. Species richness is a main component in measuring ecological diversity, so it is important to have an accurate assessment.

In both macroscopic and microscopic abundance data, the true species richness must be estimated from the observed species richness. This has been well-studied for macroscopic abundance data, resulting in, among others, an estimator by Chao (1984), a jackknife estimator by Burnham and Overton (1978), and a bootstrap estimator derived by Smith and van Belle (1984). These estimators have the common goal of increasing (extrapolating) the observed species richness in order to predict the true species richness. These estimators are ill-suited for OTU abundance data, since the observed species richness needs to be reduced rather than enlarged. Dickie (2010) applied the Chao and two versions of the jackknife estimators to simulated OTU data and found that none were adequate, and the Chao estimator was particularly sensitive to singletons.

The presence of singletons in OTU data create exceptional difficulties in estimating species richness, since a singleton could represent a unique and rare species or it could be the result of sequencing error. Unterseher *et al*. (2011) state that approximately 75% of all Roche 454 pyrosequencing singletons are 'technical artefacts' and recommend removing all the singletons from a data set prior to statistical analysis. Dickie (2010), considers this approach to be conservative, since this may eliminate some real species. Reeder and Knight (2010) developed an algorithm called Pyronoise to filter the 'true' singletons from sequencing artefacts. To reduce the number of singletons in a data set, Kunin *et al*. (2010b) recommend setting the cluster similarity threshold no greater than 97%. Jumpponen and Jones (2009) set the threshold

slightly lower, claiming that the number of OTUs is "relatively stable up to 95%", but the number of OTUs "assumes near-exponential growth at thresholds of more than 95%".

Rarefaction analysis is another method for generating species richness values and is used for both macroscopic and OTU data. Rarefaction adjusts species richness values for samples with a large number of individuals so that they can be compared to smaller samples. This is done because samples with a larger number of individuals are likely to have a larger number of unique species. The results of rarefaction analysis are usually presented in the form of rarefaction curves, as shown in Figure 2.1. There is one curve for each sample (site). To create one of these curves, the individuals observed at the site are statistically resampled (without replacement) to determine the relationship between sampling effort and number of species. For OTU data, the sampling effort is measured by the number of individuals (*i.e.* the number of sequences). To illustrate this procedure, suppose a site has 100 unique OTUs with a total count of 1400 sequences. To rarefy this site to a total count of 1000, repeated subsamples of size 1000 are selected from the 1400, and the species richness (number of distinct OTUs) in each subsample is recorded. The average species richness value is used to generate the rarefaction curve. This value can be compared to the rarefied species richness values for each of the sites that originally contained 1000 or more observed sequences. To compare a site that has only 500 observed sequences, all larger sites would need to be rarefied to 500 sequences.

Magurran (2004, page 79) warns against using rarefaction curves to extrapolate species richness, stating "The purpose of rarefaction is to make direct comparisons amongst communities on the basis of number of individuals in the smallest sample." Roesch *et al.* (2007) seem to disagree. They fit a Michaelis-Menten equation (*cf.* Graham Dunn, Encyclopedia of Biostatistics, 2005) to each curve and estimated species richness as the upper asymptote as the number of sequences increases.

Rarefaction curves can be greatly affected by the clustering thresholds used when defining the OTUs. Recall that each OTU represents a collection of detected DNA sequences that are clustered together based on their similarity and that the researcher defines clustering thresholds that control the degree of clustering. For example, if the clustering threshold is 100% similarity (or 0% dissimilarity), only those sequences that are perfect matches will be clustered. This results in a larger number of OTUs, but the abundances for the OTUs will be smaller. If the

clustering threshold is 95% similarity (or 5% dissimilarity), more clusters will be created resulting in fewer OTUs with higher abundances.

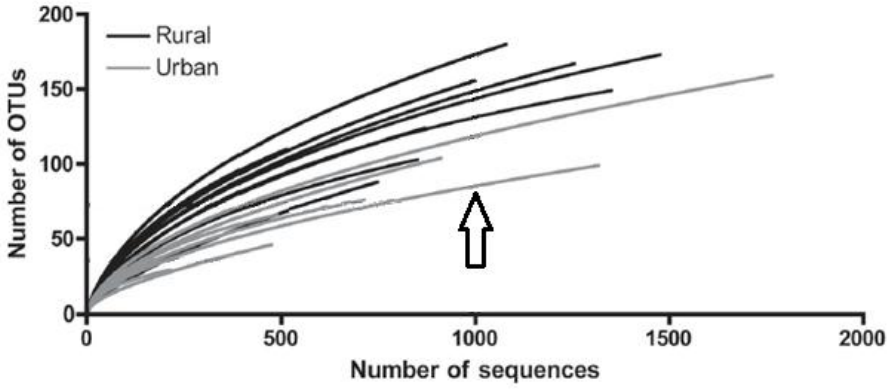


Figure 2.1: Rarefaction Curves

Each curve represents one site (one sample). The curve indicated by the arrow contains approximately 1400 observed sequences (total abundance), and species richness (unique OTUs) of approximately 100. The position of the arrow indicates that if only 1000 sequences had been observed at this site, the species richness is expected to be approximately 80. This value is obtained by repeatedly sampling 1000 sequences from the original 1400 and calculating the average species richness across all subsamples. Image adapted from Jumpponen and Jones (2009).

Roesch *et al.* (2007) provide a comparison of rarefaction curves for bacterial OTUs using percent similarity thresholds from 100% to 80% (dissimilarity thresholds from 0% to 20%). One of their comparisons, shown in Figure 2.2, illustrates the magnitude of the change in observed species richness (number of OTUs) as the clustering threshold changes. The curves representing 20% and 10% dissimilarity are very close, and appear to have an asymptote at approximately 500. At 5% and 3% dissimilarity, the asymptotes are approximately 1800 and 2100. In contrast, the curve for 0% dissimilarity has not yet begun to level off, so the estimate of the asymptote is uncertain. These curves are all based on the same data, and could result in species richness estimates of 500 to 2100, depending on the clustering threshold.

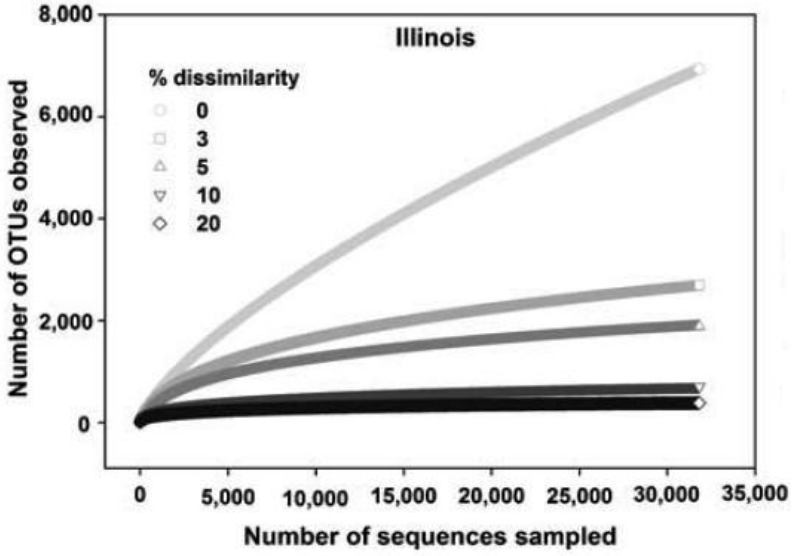


Figure 2.2: Affect of Clustering Thresholds on Rarefaction Curves

The clustering thresholds are used to determine how similar two DNA sequences need to be in order to classify them as the same Operational Taxonomic Unit (OTU). For 0% dissimilarity, the two DNA sequences need to be a perfect match. This results in a large number of distinct OTUs. Increasing the dissimilarity threshold generates fewer distinct OTUs. Image source: Roesch et al., 2007.

2.3. Defining Distance

To examine the effect of experimental and/or environmental conditions on the assemblage of species at the sites, it is necessary to have some measure of how similar or dissimilar two sites are. Since each site is characterized by a vector of OTU counts, it is natural to express the dissimilarity between sites as a form of multivariate distance. When dealing with multivariate data, defining a reliable measure of distance, or dissimilarity, has many challenges. Ordinary Euclidean distance is known to be dominated by the dimension with the largest variability, and it is difficult to specify what constitutes 'closeness' in high-dimensional data. With ecological data, it is sometimes easier to conceptualize similarity, then apply a mathematical transformation to measure dissimilarity. Ecological similarity between two sites is based on the number of shared species and abundance values they have in common.

From a mathematical perspective the term 'distance' is the same as 'metric', while a 'dissimilarity' is a more general (less precise) way of describing the separation between two

objects. The term distance is often used as a synonym for dissimilarity. Strictly speaking, a distance between x and y is any function $d(x,y)$ that satisfies four conditions:

(1) $d(x,y) \geq 0$ (non-negativity)

(2) $d(x,y) = 0$ if and only if $x = y$ (isolation)

(3) $d(x,y) = d(y,x)$ (symmetry)

(4) $d(x,y) \leq d(x,z) + d(z,y)$ (triangle inequality)

A function that satisfies only the first three conditions (and not the fourth) is considered a dissimilarity, although we will use the terms distance and dissimilarity interchangeably. The triangle inequality is an important criterion if distances are to be used in clustering (partitioning the objects according to similar characteristics) or ordination (positioning the objects in a space that contains fewer dimensions than in the original data set). Although both of these operations can use dissimilarities rather than distances, the results can be less reliable with dissimilarities.

When we visualize the abundance data as a matrix, with rows representing sites and columns representing species, dissimilarity can be measured as a distance between rows (Q mode analysis) or as a distance between columns (R mode analysis). Furthermore, the distance can be based on absolute abundance, relative abundance or absence/presence indicators. In general, quantitative measures (based on abundance) are generally superior to qualitative measures (based on absence/presence), although quantitative measures "can be unduly influenced by the abundance of the most dominant species" (Magurran, 2003, p. 175). To alleviate this problem, Clarke and Warwick (2001) recommend transforming the data via square roots or logarithms.

To quantify the amount of separation between sites (rows) or species (columns), we treat the rows (or columns) as vectors and combine the separation between each pair of elements. Metrics include ordinary Euclidean distance (the L_2 norm) and the Manhattan or city block distance (the L_1 norm). Many of the existing dissimilarity measures for ecological data are variants of one of these two norms. Other measures have been proposed based on, for example, chord distance, chi-squared distance, and probabilistic measures such as Kullback-Leibler distance. These types of measures have not been readily accepted by the ecological community because, in part, they fail to capture the unique characteristics of ecological distance. Faith, Minchin and Belbin (1987) performed an extensive comparison of dissimilarity measures for

ecological data and concluded that the Kulczynski, Bray-Curtis and Relativized Manhattan measures were the most robust in terms of maintaining a linear, rank-order relationship with the corresponding distances in ecological space.

Similarity and dissimilarity are at opposite ends of a continuum of values that represent the degree of association. While some measures are defined as similarity measures and others are formulated as dissimilarity, every similarity measure can be transformed to become a dissimilarity measure, and vice versa. Similarity measures S are generally defined to have range [0, 1], so the corresponding dissimilarity measure D can be defined as $D = 1 - S$, $D = \sqrt{1-S}$ or $D = \sqrt{1-S^2}$. If a dissimilarity measure is constrained to the unit interval, then a comparable transformation can be used to obtain similarity. If, however, the dissimilarity is not bounded (such as a true distance), then it may be possible to constrain the distance via standardization.

Magurran (2004, p. 174) specifies six desirable criteria for a similarity index between sites, but notes that very few existing indices satisfy all six. The criteria are:

(1) the value should be 1 (or 100) when two samples are identical
(2) the value should be 0 when samples have no species in common
(3) a change of measurement unit does not affect the value of the index
(4) the value is unchanged by the inclusion or exclusion of a species that occurs in neither sample
(5) the inclusion of a third sample makes no difference to the similarity of the initial pair of samples
(6) the index reflects differences in total abundance (and not just relative abundance).

Numerous measures for similarity and dissimilarity have been proposed for ecological data, which makes comparison cumbersome. This is complicated by the fact that many measures have been formulated independently by different researchers, and are therefore assigned different names. In addition, the name can change depending on whether the measure uses absolute abundance, relative abundance or absence/presence indicators. For example, one of the most widely used dissimilarity measures is the Bray-Curtis distance, which is also known as the quantitative Sørensen index and the Czekanowski distance (Cha, 2007).

The R Package (R Development Core Team, 2009), specifically the library *vegan*, (Dixon, 2003) has 14 built-in dissimilarity measures, and also provides a program in which users can define their own dissimilarity function. Exploring all of these is beyond the scope of the current paper. Instead we focus on three: Bray-Curtis, Kulczynski and Morisita-Horn. These

measures were chosen because they are particularly robust (Faith *et al.*, 1987; and Magurran, 2004, page 174) and because they are all restricted to the interval [0, 1] so that direct comparison is possible. These distance measures, as applied to distances between sites, are defined below. To obtain the distance between OTUs, simply take the transpose of abundance data matrix.

Let x_{ij} represent the abundance of OTU j at site i and let T_i represent the total abundance at site i.

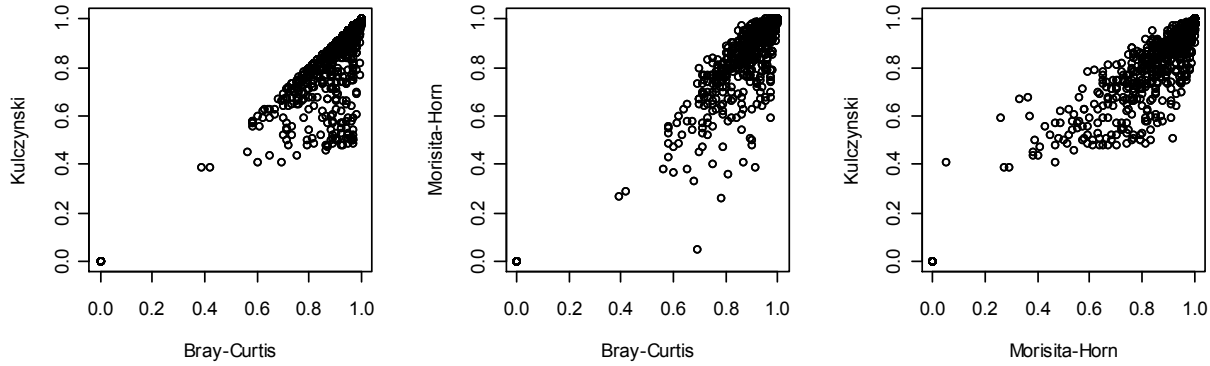
Bray-Curtis: $$\mathrm{BC}_{ik} = \frac{\sum_j \left| x_{ij} - x_{kj} \right|}{T_i + T_k}$$

Kulczynski: $$\mathrm{KUL}_{ik} = 1 - \frac{1}{2}\left\{ \frac{\sum_j \min\left(x_{ij}, x_{kj}\right)}{T_i} + \frac{\sum_j \min\left(x_{ij}, x_{kj}\right)}{T_k} \right\}$$

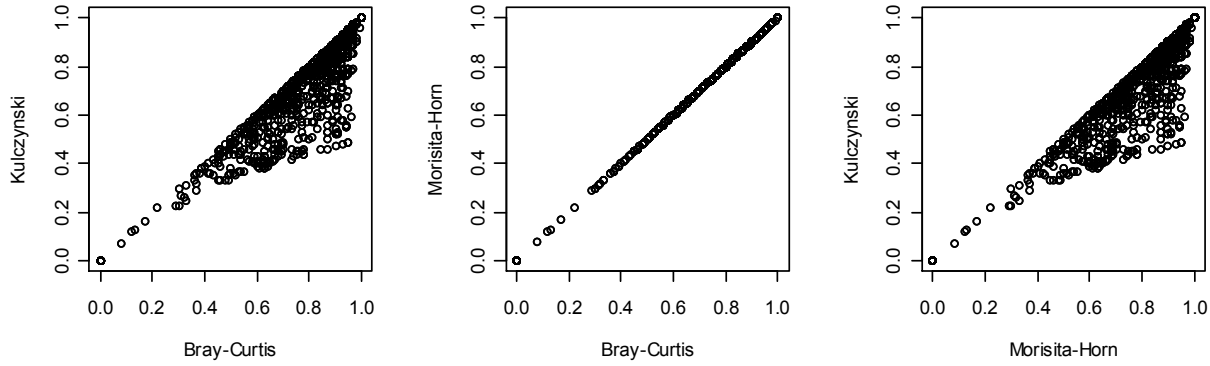
Morisita-Horn: $$\mathrm{MH}_{ik} = 1 - \frac{2\sum_j x_{ij} \cdot x_{kj}}{\left(\frac{\sum_j x_{ij}^2}{T_i^2} + \frac{\sum_j x_{kj}^2}{T_k^2} \right) * \left(T_i * T_k\right)}$$

We explore these distances using a subset of the soil data, which is described in Section 3.1. The results are shown in Figure 2.3. Note that the intent of these comparisons is to explore the differences between the distance measures and not to interpret the differences between sites. When using the actual abundance data, the Bray-Curtis distance is generally larger than both the Kulczynski and Morisita-Horn distances, but the latter two measures are comparable. When the abundance values are replaced by 0/1 indicator variables, the Bray-Curtis and Morisita-Horn measures are identical, and generate larger distances than the Kulczynski measure. The triad of graphs in (c) illustrate how each measure compares to itself, using abundance values versus absent/present indicators. For both the Bray-Curtis and Kulczynski measures, the distances are greater when they are based on the abundance values. In contrast, the Morisita-Horn measure is less influenced by the total abundance, generating similar distances for both abundance values and absent/present indicators. Whether we work with the raw abundance values or absent/present indicators, there seems to be no unambiguous method to measure distance in ecological data.

(a) Using abundance data



(b) Using absent/present (0/1) data



(c) Compare abundance to absent/present

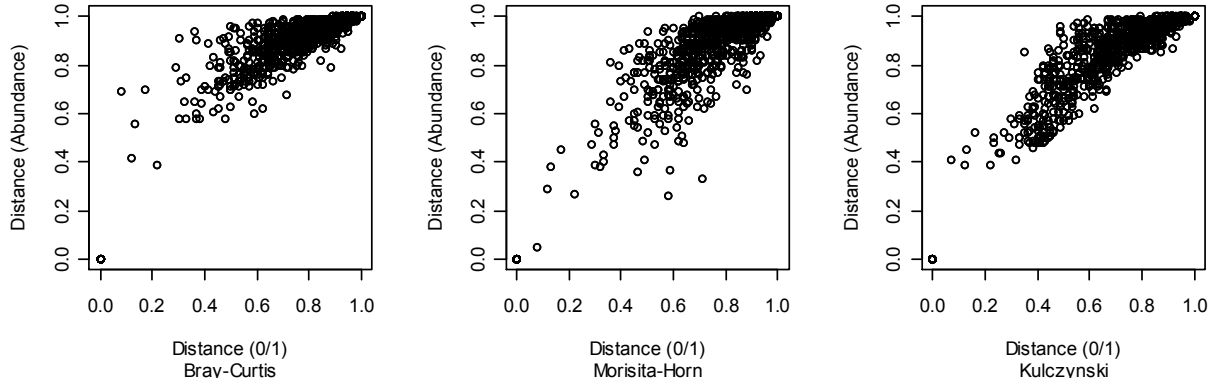


Figure 2.3: Comparison of Three Distance Measures

Distance between randomly selected OTUs from the soil data, described in Section 3.1. Each point represents the distance between one pair of OTUs.

2.4. Comparing Sites

2.4.1. Site Totals

Laboratory protocols for preparing each analyte should, in theory, generate approximately equal total counts (abundances, sequences) for each analyte (sample, site). In practice, however, there can be excessive variation in these totals, creating outliers in the distribution of site totals. This occurred, for example, in the soil data described in Chapter 3, in which soil samples taken from lower depths included larger concentrations of clay. The clay inhibited DNA extraction from these samples, resulting in extremely low total counts.

Sites with larger total abundance are likely to have larger values for species richness, which can affect measures of diversity. From personal conversations with Dr. Ari Jumpponen and Dr. Karen Garrett, it appears that sites with extremely low totals are simply discarded. Most published articles do not mention outliers among the site totals, although Jumpponen, *et al.* (2009, 2010b) use an ANOVA F test to confirm no significant differences in mean site totals across the experimental factors. Another approach is presented in Ishak *et al.* (2011), who report no specific test for equality of site totals, but state that "Because the numbers of sequences acquired from different samples varied substantially, we randomly selected 1,000 sequences from each sample" (page 823). Subsequent analysis in Ishtak *et al.* was based on the data obtained from the single resampling event.

2.4.2. Measures of Diversity

There are many types of diversity, but we consider only two: within-site (alpha) diversity and between-site (beta) diversity. Alpha diversity is often used as a univariate measure to test for differences between sites. That is, a null hypothesis of no difference between sites can be formally stated as: There are equal values of alpha diversity across the sites. See, for example, Gao and Yang (2010) and Van Diepen *et al.* (2011). Beta diversity is a measure of dissimilarity between sites, and is most often used as a surrogate for distance. There are many different ways to measure both alpha and beta diversity, and these measures differ in the emphasis they place on richness as opposed to evenness. Magurran (2004, page 101) advises that, while is tempting to use a variety of diversity measures and compare the results, this is not a good practice. Instead,

she lists nine key points to consider in selecting an appropriate measure. Some of the most widely used measured are described below.

2.4.2.1. Alpha (within-site) Diversity

For each site, α diversity is a combination of the site's species richness and species evenness. Three of the most common methods of measuring α diversity are the Shannon index, Simpson's D, and Fisher's α. To define these measures, let S represent the number of distinct OTUs detected at the site, and let x_i represent the abundance (count) for OTU *i*. Then $N = \sum_{i=1}^{S} x_i$ is the total count (number of sequences) for the site and $p_i = \frac{x_i}{N}$ is the sample proportion for OTU *i*. These calculations exclude all counts that are zero. **Shannon's index**, sometimes erroneously called the Shannon-Weaver index, is defined by $H = -\sum_{i=1}^{S} p_i \log_b p_i$. If the logarithmic base is *e*, the index is called H'. The term "Shannon's index" can also refer to Shannon's evenness index, which is defined by $E = H' / \ln S$. **Simpson's D** estimates the probability that two individuals drawn at random will not belong to the same OTU. For infinite populations, it is defined by $D = 1 - \sum_{i=1}^{S} p_i^2$, but when estimating this from a sample it is customary to use the formula for finite populations: $D = 1 - \sum_{i=1}^{S} \frac{x_i(x_i - 1)}{N(N-1)}$. The term "Simpson's D" sometimes refers to similarity rather than diversity. The similarity is defined to be $1 - D$. **Fisher's** α is based on the log series distribution (see Appendix B) and is sometimes called log series α. The log series is defined by $\frac{\alpha\theta^k}{k}, k = 1, 2, \ldots$, where the k^{th} term in the series is the number of OTUs predicted to have exactly *k* individuals. If each term in the series is divided by S (the number of unique OTUs in the sample), then this series becomes the log series probability distribution. Since the discrete probability distribution must sum to 1, log series must sum to S, so the value for α is defined by $\alpha = \frac{-S}{\ln(1-\theta)}$. Thus θ is the only parameter of the series. Fisher's α is the maximum likelihood estimate of α. Magurran (2004, page 30) describes

how to estimate Fisher's α: First iteratively solve $\frac{S}{N} = \frac{(\theta - 1)}{\theta} \cdot \ln(1-\theta)$ to get $\hat{\theta}$, then $\hat{\alpha} = \frac{N\left(1-\hat{\theta}\right)}{\hat{\theta}}$. This result is verified in Appendix B.

2.4.2.2. Beta (between-site) diversity

Beta diversity measures the amount of separation between two sites, which can be interpreted as the 'distance' between the sites. Thus distance measures such as those defined in Section 2.3 are used as measures of beta diversity. Empirical estimates of the various beta diversity measures "are often uncorrelated and can provide different but equally illuminating views of diversity" (Lozupone, *et al.,* 2007, page 1576). Diversity is a measure that combines species evenness and species richness, and the various beta diversity measures differ in the degree to which they emphasize evenness versus richness.

2.4.2.3. Absent/present vs. abundance

If the observed data are recorded as binary (absent/present indicators) instead of the actual abundance (counts), then α diversity is measured by species richness, the number of distinct OTUs at the site. The binary measures for beta diversity are, in most cases, derived from the abundance-based beta diversity measures, but the formulas simplify dramatically and the name of the measure can change. The binary measures are usually defined in terms of the elements of a 2x2 contingency table, where *A* is the number of OTUs that are present at both sites, *B* is the number of OTUs present only at the first site, and *C* is the number of OTUs present only at the second site. The fourth element in the contingency table (the number of OTUs that are not present at either site) is generally not used in these calculations, since it provides little information about the dissimilarity between the sites. One common index is the **Jaccard index**, also called the Marczewski-Steinhaus distance, and is defined by $1 - \frac{A}{A+B+C}$ or $\frac{B+C}{A+B+C}$. Another popular index is the **Sørensen index** (or Dice coefficient), defined by $\frac{2A}{2A+B+C}$, which is identical to the Bray-Curtis measure as applied to binary data.

2.4.3. Testing for Differences between Sites

Most ecological applications for OTU data involve comparing sites, where the sites can be separated by space (*e.g.* urban vs. rural, Jumpponen and Jones, 2009, 2010a) or separated by time (*e.g.* glacial retreat, Fujiyoshi *et al.*, 2011). Statistical tests for detecting differences between sites can be based on multivariate techniques that utilize the entire vector of OTU counts observed at each site, or can be univariate techniques applied to any one of the diversity measures calculated for each site. Univariate techniques can also be applied to each OTU separately. To compare experimental and/or environmental conditions, the sites are grouped according to these conditions. Univariate tests across multiple conditions typically employ ANOVA or the nonparametric Kruskal-Wallis test, and two-condition comparisons often utilize the Mann-Whitney test or the t-test. The normality conditions required by both ANOVA and t-tests are satisfied because "the Shannon, Simpson, and other widely used diversity statistics are often approximately normally distributed" (Magurran, 2004, page 151). In two separate studies of fungal OTU data, van Diepen *et al.* (2011) employed two-way ANOVA to compare the effects of nitrogen treatments and Gao and Yang (2010) performed a Kruskal-Wallis test to detect differences across two experimental factors. Both of these analyses used Shannon's diversity index. As another example, Ishak *et al*. (2011) used a species richness estimator and a Mann-Whitney test to compare bacterial OTU communities in ant colonies.

If the difference between sites is tested using the entire vector of OTU abundances at each site, then a permutation test can be used. In these tests, groups of sites are labeled according to the experimental and/or environmental conditions, group labels are permuted among the sites, and an F-type test statistic is used to compare within-group variability to between-group variability. For one-way designs, the Multiple Response Permutation Procedure (MRPP) can be used (Zimmerman, 1985) . This procedure is available in many statistical software systems, including SAS, SPSS, and in the R package *vegan*. Permutation tests for two-way designs, including possible interaction terms, can be done with PERMANOVA (Anderson 2001, 2005), although this requires that the experimental design be balanced. PERMANOVA is implemented as a standalone FORTRAN program, available as a free download from http://www.stat.auckland.ac.nz/~mja/Programs.htm.

2.5. Describing Species: Common vs. Rare

Species (OTUs) can be characterized by their prevalence (the number of sites at which they occur) and by their pattern of abundances at these sites. Commonly occurring species, also called resident or core species, often have different patterns of abundance than rare, or satellite, species. For macroscopic species, Magurran and Henderson (2003) propose a method to classify species as either core or satellite. Their method uses a 50% persistence threshold, that is, species that occur in at least 50% of the sites are classified as core species and those that occur at less than half the sites are satellite species. The abundances of core species, when viewed across all samples, are usually modeled with a lognormal distribution, while the abundances of satellite species are modeled with a log series distribution. Both groups have highly skewed distributions, but the satellite species tend to be more skewed.

Unterseher *et al.* (2011) applied the 50% persistence threshold to three fungal OTU data sets. To assess the effectiveness of this procedure, all species were compared to the lognormal distribution, then the species were split into the two groups and each group was compared to the lognormal distribution. For each of the three groups (core, satellite, and combined) and each of the three data sets, the goodness of fit for log-normality was assessed via the chi-square, Anderson-Darling, Kolgmorov-Smirnov, and Shapiro-Wilk tests. For every data set and every testing method, the p-value for the core group was larger than the p-value for the combined group, indicating that the core group is more likely to follow a lognormal distribution than the two groups combined.

In a completely different approach, Scott T. Bates and colleagues at the University of Colorado, Boulder (unpublished work) are developing methods for creating networks to illustrate the relationships between OTUs. These networks use only the most abundant OTUs, and the degree of association between each pair of OTUs is measured by Pearson's correlation. However, the use of a correlation coefficient is not recommended by Legendre and Legendre (1998, page 293), since it measures only a linear relationship and will fail to detect two species that always occur together, but "do not covary in a linear way."

Christopher J. van der Gast, *et al.* (2011) used the index of dispersion as a test statistic for categorizing bacterial OTUs from lung tissue sampled from cystic fibrosis patients. The index of dispersion is defined as the ratio of the variance to the mean. For a random sample of size n from a Poisson distribution, this index is approximately distributed as chi-square with $n - 1$

degrees of freedom (Selby, 1965). van der Gast's procedure is based on the assumption that rare OTUs are "randomly distributed through space" (page 785) and therefore the individual abundances for these OTUs follow a Poisson distribution. Each OTU was tested separately. If the index of dispersion for an OTU fell outside the 95% confidence limits for a $\chi^2(n-1)$ distribution, the OTU was classified as core. Otherwise, the OTU was classified as rare. The adequacy of the classification was assessed using both the χ^2 and Kolmogorov-Smirnov goodness-of-fit tests. Within the core group, the collection of OTU total abundances was compared to a lognormal distribution. Within the rare group, the collection of OTU total abundances was compared to a log series distribution. For both groups and for both tests, no significant deviation from the target distribution was detected. When the two groups were combined, neither the lognormal nor the log series distribution fit the data. It was therefore concluded that the classifications were accurate.

The cystic fibrosis data set consists of 82 OTUs measured on 14 patients and the total abundance (for all OTUs and all patients) is 2139. The hypothesis test utilizing the index of dispersion divided the OTUs into 15 core and 57 rare OTUs. The relatively small number of core OTUs account for 89.9% of the total abundance, while the large number of rare OTUs account for only 11.1% of the total abundance. These percentages are typical for OTU data sets, but the dimension of the cystic fibrosis data set is considerably smaller than the four ecological data sets described in Section 3.1. For the cystic fibrosis data, the index of dispersion tests were based on a chi-square distribution with 13 degrees of freedom, resulting from observations on 14 patients. For the soil data (described in Section 3.1), these tests would have 237 degrees of freedom based on observations from 238 sites. Such a large value for degrees of freedom will give this test high power and increase the likelihood of detecting very small departures from the hypothesized distribution, departures that are not of practical significance. For this reason, further investigation involving the index of dispersion has not been pursued at this time, but could be considered in the future when comparing methods for classification into rare versus core species.

2.6. Probability Models

There are many ways to view the natural variability that occurs in ecological data sets. If we envision the data as a matrix, with rows representing sites and columns representing species (or OTUs), then each entry in the matrix is the number of individuals (abundance) of one species observed at one site. There is natural variability among the entries of the matrix, which give rise to variability in the row totals and column totals. Variability in row (site) totals is often related to sampling effort, in that sites subjected to low sampling effort are more likely to have lower site totals than those sites subjected to greater sampling effort. As discussed in Section 2.4.1, site totals for OTU data sets are expected to be approximately equal, and sites that have 'unusually' low totals are usually removed from the data prior to analysis. The literature review has uncovered no discussion regarding an appropriate probability distribution for the site totals, and the decision to remove a site from the data set is usually made on a subjective basis.

In contrast, the literature is rife with examples of probability models for a collection of species within a community. These are called Species Abundance Distributions, or SADs, and are one of the most basic descriptions of an ecological community. In this sense, a 'community' can be a single site, a collection of similar sites, or all the sites in the data set. An SAD describes the number (or proportion) of species in the community predicted to have a particular abundance. In some cases, the abundance values are log-transformed and binned into octaves, with probabilities assigned to octaves rather than to raw abundance values (see Section 3.1.3). When based on empirical data, SADs can be represented in a variety of ways, including histograms, rank-abundance diagrams (*aka* Whittaker plots), or cumulative distribution functions. When raw abundance values are plotted as a histogram, "every community shows a hollow curve ... with many rare species and just a few common species" (McGill *et al.*, 2007). The formation of the hollow curve shape is considered to be one of the universal laws of community ecology.

Although the shape of the histogram is undisputed, there seems to be no clear consensus regarding how to mathematically describe this distribution. A wide variety of probability models have been proposed, but these are based on observed abundance patterns in macroscopic organisms and it is unclear if these are appropriate for OTU data. Some of the more popular models are:

- the geometric distribution, which predicts extremely uneven abundances;

- the broken stick distribution (MacArthur, 1957), which predicts extremely even abundances;
- the log series distribution, which predicts a high proportion of very rare species;
- the lognormal distribution, which predicts a low proportion of very rare species.

The confusion regarding an appropriate distribution should be apparent in the fact that the first three of these distributions are discrete, while the fourth is continuous. While the lognormal distribution is almost universally accepted as an appropriate distribution for common species (see, for example, Magurran, 2004; Ludwig, 1988), Williamson and Gaston (2005) and Williamson (2010) claim that the lognormal distribution is *never* appropriate for *any* species abundance distribution since the fundamental assumptions underlying the lognormal distribution are never satisfied. Numerous other distributions have been proposed, and some have been specifically created, to describe SADs. McGill *et al.* (2007) list such 27 models, but acknowledge that this is a partial list. Confusion is exacerbated when we consider species that are measured rather than counted. This occurs, for example, in vegetation studies where the species are measured in terms of their biomass.

2.7. Reducing Dimensionality

OTU data sets typically contains thousand of OTUs, but many of these are rare OTUs, that occur in very few samples with very low total abundance. In order to visualize relationships between OTUs, it is often beneficial to reduce the dimension of these data sets. Some researchers simply remove rare OTUs and others combine OTUs into higher-order taxonomic groups. Jumpponen and Jones (2010a) employed both strategies. In their first analysis, "OTUs that occurred in more than 20% of the samples ... and were represented by more than 100 [total count] were analyzed ..." In a second analysis of the same data, each "nonsingleton OTU was assigned to a genus, family and order based on BLAST matches" and within each site the OTU abundances were summed to create genus-level abundances. (BLAST was developed by Zhang *et al.* (2000) and is part of a collection of public databases and associated bioinformatics tools for genetics researchers maintained by the National Center for Biotechnology Information, available at http://blast.ncbi.nlm.nih.gov/Blast.cgi.)

There are other methods for reducing the dimensionality of OTU datasets, including principal component analysis (PCA), principal coordinate analysis (PCoA) and nonmetric

multidimensional scaling (NMS). In a data set with S OTUs , each vector of site abundances can be represented in S-dimensional real space, $\Re^S$. These three methods rotate and re-scale this space so that the vectors can be expressed in a reduced space, while preserving the spatial relationship between the vectors (the distance between the sites). This facilitates comparison of the sites, and may uncover relationships that are obscured in the full space $\Re^S$.

PCA is a common method in other statistical applications but is rarely used with ecological data. As Legendre and Legendre (1998, page 292) explain, species abundance data is characterized by many zeros, which distort the dispersion (or correlation) matrix upon which PCA relies. In addition, the first few principal components are strongly influenced by the extreme skewness in the data, presenting a distorted view of the data in reduced space. This is partly because PCA utilizes ordinary Euclidean distance, which is unsuitable for species abundance data.

Principal coordinate analysis (PCoA) is similar to PCA but can operate on any dissimilarity measure, including ones that are not metrics (do not satisfy the triangle inequality). While this is an improvement over PCA, it is not as effective as nonmetric multidimensional scaling (NMS) at "compressing distances relationships among objects into, say, two or three dimensions." (Legendre and Legendre, 1998, page 425). Like PCoA, NMS can use any dissimilarity matrix and its main purpose is to represent the data in a reduced dimensions while preserving the distance relationships between the sites. The difference between PCoA and NMS is the stringency with which they preserve the original distances. When all of the PCoA space is used, the distances (or dissimilarities) between OTUs in PCoA space are precisely the same as they were in the original space. In contrast, NMS preserves only the *order* of the distances, so that the smallest distance in the original data is still the smallest distance in NMS space, but the numeric values of these distances may be different. For most ecological applications, preserving the ordering of the distances is sufficient. Both NMS and PCoA are widely used as methods of dimension reduction, for example, by Geml *et al.* (2010), Jumpponen *et al.* (2010b), Ishak *et al.* (2011) and Lozupone *et al.* (2007).

Chapter 3. Exploratory Studies

This chapter describes preliminary results by the author of this proposal, and it builds off of an initial collaboration that resulted in the paper by Jumpponen, Keating, Gadbury, Jones, and Mattox (2010). In fact, results in that paper included exploratory analyses conducted by this researcher that highlighted some of the complexities in pyrosequencing data. Section 3.1 covers exploratory analysis of four pyrosequencing data sets to determine common characteristics that can highlight needs and challenges for statistical methods development. The results of this section help to motivate the methods presented in Chapter 4.

Section 3.2 examines compositional data analysis in the context of pyrosequencing data, and it argues that one will likely gain little if compositional data analysis is used on such data and lose little, if anything, from not doing such analysis. (A review of compositional data analysis is given in Appendix D.) Section 3.3 introduces two new statistics that could be used to classify rare versus common OTUs. These statistics would replace the 50% persistence measure customarily used in macroscopic data. Later, in Section 4.3, we provide a rigorous method that generates a probability-based assignment of an OTU into a rare versus common category. Issues related to the choice of appropriate probability models are discussed in Section 3.4. Such choices are a necessary first step for the development of valid statistical procedures to test for commonalities and differences in OTU abundances, and for detecting outliers that possibly result from the data processing technology.

3.1. Comparison of Four Data Sets

Four OTU data sets are compared for the purpose of identifying the characteristics of this type of data. Of particular interest is identifying anomalies in the data, in order to explore methods to mitigate the anomalies. We are also interested in characterizing the distribution of these data, in order to identify appropriate statistical tests and to realistically simulate the data. Two of these data sets are from Dr. Ari Jumpponen and two are from two different students from Dr. Karen Garrett's workshop the last week of July, 2011. Table 3.1 provides some summary of the four data sets.

The Four Data Sets

AJ Leaf Data: Dr. Jumpponen has performed at least two distinct experiments in which fungal DNA was extracted from the leaves of bur oak trees. The first experiment is documented in Jumpponen and Jones (2009) and the second in Jumpponen and Jones (2010a). We examine the data from the second experiment. There are 34 samples and 598 OTUs. For the individual counts, 87.1% are zeros, and among the nonzeros, 25.2% are singletons.

AJ Soil Data: This data set is also from Dr. Jumpponen. Fungal DNA was extracted from soil samples at five different depths (10, 20, 40, 60 and 100 centimeters). The depth 100 samples did not generate sufficient DNA to be included in the analysis. Excluding the depth 100 samples, there are 238 samples and 2,422 OTUs. For the individual counts, 96.3% are zeros and, among the nonzeros, 48.6% are singletons.

Lorena's Data: These data were generated by Lorena Gomez Montano, one of Dr. Karen Garrett's students. This is fungal DNA from soil samples. There are 37 samples and 799 OTUs. Among the individual counts, 91.3% are zeros and, of the nonzeros, 56.2% are singletons.

Neshmi's Data: These data were generated by Neshmi Salaues Mendoza, one of the off-campus participants in Dr. Garrett's workshop. It is bacterial DNA with 30 samples and 21,620 OTUs. Among the individual counts, there are 91.3% zeros and, of the nonzeros, 63.0% are singletons.

Data Set	AJ Leaf	AJ Soil	Lorena	Neshmi
Number of Samples	34	238	37	30
Number of OTUs	598	2,422	799	21,620
Number of Zeros	17,704	554,971	26,985	591,988
Percent Zeros	87.1	96.3	91.3	91.3
Number of Individual Singletons	662	10,442	1,450	35,681
Percent of Nonzeros that are Singletons	25.2	48.6	56.2	63.0
Total Count	46,179	154,195	17,428	154,112

Table 3.1: Summary of Four Data Sets

3.1.1. Site Totals

According to Dr. Garrett (personal communication, July 2011), there should be approximately equal numbers (OTU counts) in each sample, and that excessive variation in these counts may indicate problematic DNA extraction and/or PCR amplification. The histograms in Figure 3.1 show the wide variation of sample (site) totals in each data set. Lorena's data indicate one site with a distinctly low total, but the remaining site totals appear approximately symmetrically distributed. Neshmi's data set shows a slightly skewed distribution, and contains what could be one small and 2 (or perhaps 4) large sites. The leaf data are not symmetric, but there do not appear to be any extreme outliers. The soil data are unique in both the range of site totals and the handful of extremely large values.

It is fairly clear, even without any statistical tests, that the one small sample in Lorena's data is inconsistent with the remaining site totals in that data set. Lorena has indicated that this sample will be discarded. A similar situation occurs in soil data, but in this data set a few 'overly' large site totals appear to be incompatible with the rest. Unlike Lorena's data, the soil data contain no clear separation between acceptable and unacceptable site totals, thus some form of systematic investigation is warranted. For both Neshmi's data and the leaf data, there appear to be obvious signs of outliers.

To better assess the differences in the empirical distributions of the site totals, particularly the small site totals, we now consider a logarithmic transformation. The histograms are shown in Figure 3.2. Lorena's data, excluding the one small sample, are similar in shape to the leaf data. Neshmi's data show two 'large' samples, although this interpretation is subjective. The soil data are unique due to its numerous small samples, which were not apparent until the data were log-transformed. For the soil data, the median sample count is 606, but there are 12 sites (out of 238) that each have total count less than 100, and the smallest of these is only 7. Sites with such small totals cannot be directly compared to a site with 600 sequences. Samples that are too small are considered not viable and should be removed from the data set, but it is currently unclear at what point a site total is too small to be considered viable. These determinations are not regularly reported in published articles, although Jumpponen and Jones (2010a) and Jumpponen *et al.* (2010b) report using one-way ANOVA to test for equitable site totals across experimental conditions. In both cases, no significant differences in site totals were detected.

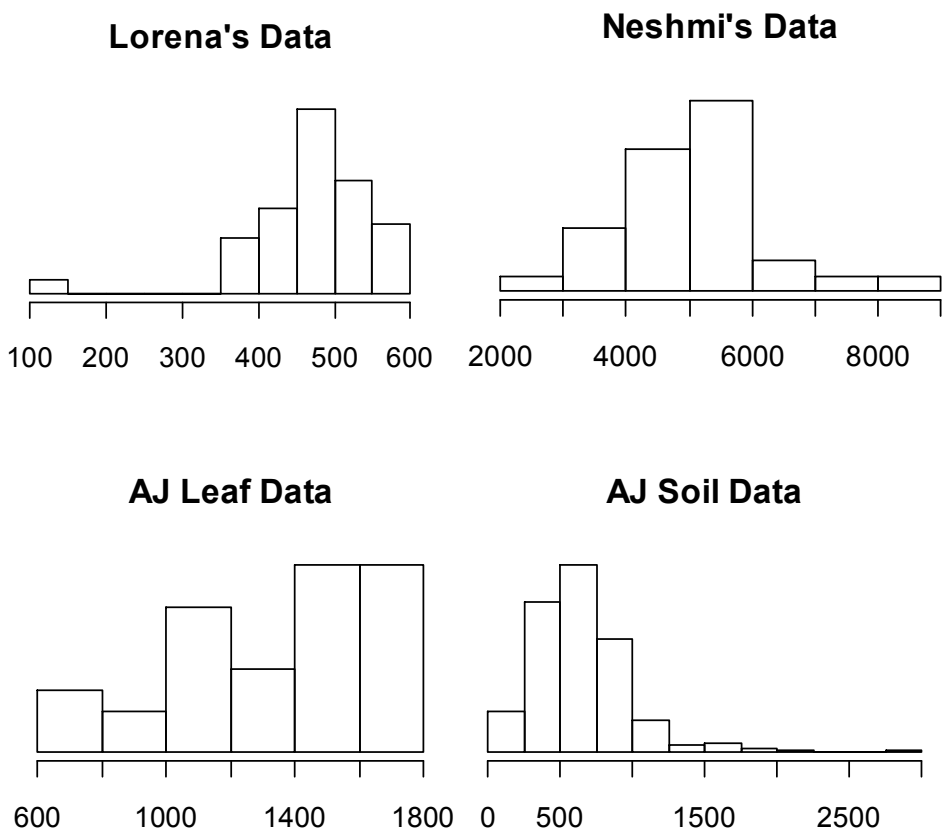


Figure 3.1: Distribution of Site Totals for Four OTU Data Sets

Site totals should be approximately equal, and extreme imbalances may indicate problematic DNA extraction or amplification. Sites with small site totals are usually discarded. Sites with large site totals can be rarefied for comparison to sites with smaller totals. See the text for details on rarefaction.

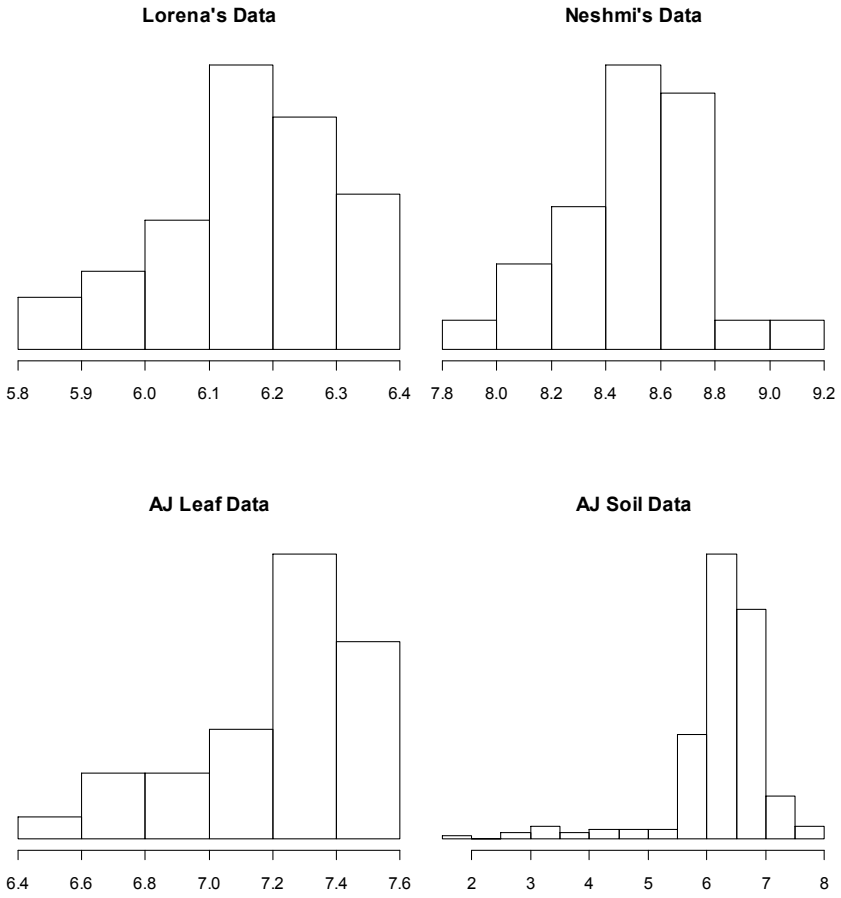


Figure 3.2: Distribution of the Logarithm of Site Totals

A logarithmic transformation reveals different patterns in the distribution of site totals. In the soil data, for example, there are many sites with extremely small totals. This pattern is not evident until the data are transformed.

As shown in the previous histograms, site totals that are too small may not be clearly distinguished when examining the raw site totals. A logarithmic transformation is well suited to separate the small values, but this will tend to obscure unreasonably large values. Rather than binning the logarithmic site totals to create a histogram, it may be beneficial to view these values in comparison to the number of unique OTUs in the site. Graphs for the four data sets are shown in Figure 3.3. Lorena's single small sample is clearly shown an unusual point, while neither Neshmi's data nor the leaf data indicate unusual site totals. The soil data, in contrast, show many 'suspicious' points, as indicated by the long tail to the left. It is possible that these small samples occur by chance alone, since the soil data contain approximately six times the number of samples

as the other data sets. Note that the logarithmic scale obscures the large site totals in the soil data.

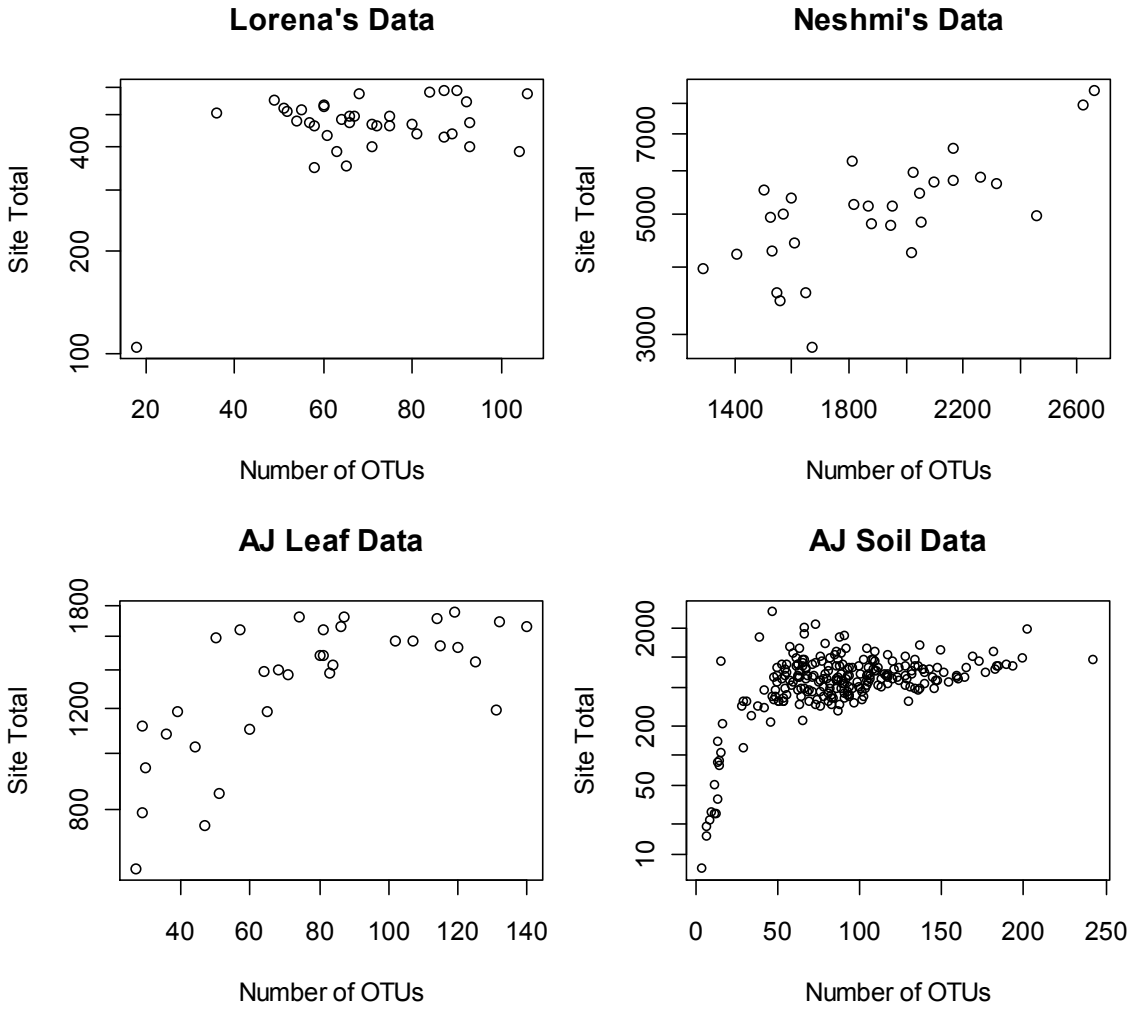


Figure 3.3: Species Richness (Number of OTUs) and Site Abundance

Each point represents one site (one sample), and shows the number of distinct OTUs observed at the site versus the total abundance for the site. Site totals are given in logarithmic scale.

We can also use side-by-side boxplots to compare the distributions of site totals in these four data sets. The measures of center (the median) for these four data sets are quite different, from a low of 475 in Lorena's data to a high of 5089 in Neshmi's data. To facilitate comparison, we standardize the values within each data set by subtracting the median and dividing by the interquartile range. The distributions of the standardized totals are shown in Figure 3.4. Lorena's single small sample is clearly shown as an outlier, as are two large samples in Neshmi's

data. Note that this plot completely obscures the numerous small samples in the soil data, most likely because this data set contains both large and small samples, and the small values are truncated at zero while the large samples have no upper bound.

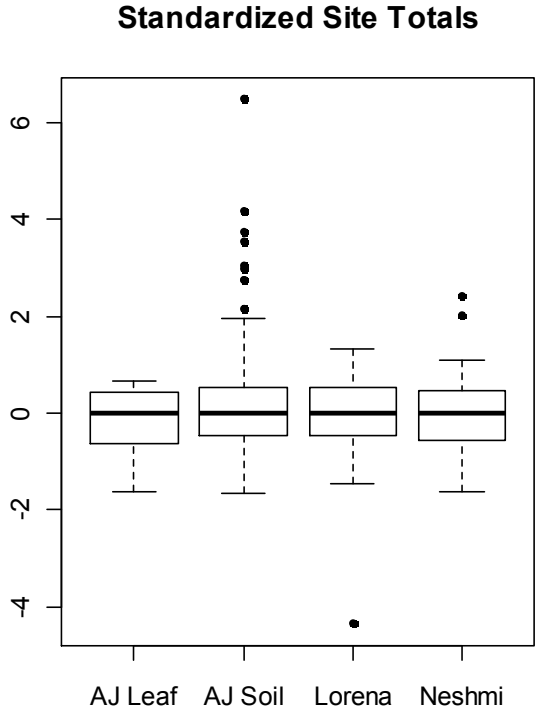


Figure 3.4: Standardized Site Totals

For each data set, the total abundance for each site is standardized by subtracting the median and dividing by the interquartile range.

3.1.2. Detecting Outliers in Site Totals

Although it is recognized that extreme variations in site totals is undesirable, there appears to be no obvious method for detecting outliers in these totals. Graphs such as those in the preceding subsection can assist in identifying potential outliers, but they provide no statistical basis for identifying a site total as an outlier. It is known that variations in PCR amplification can have a direct impact on site totals, but it is unclear if there is a theoretical basis for an underlying probability distribution for these totals, which would give rise to a parametric statistical testing procedure. Currently, sites with low totals are deemed improperly amplified and are simply discarded, and sites with overly large totals are rarefied in order to estimate species richness. It is unclear if any other accommodations are made for large site totals. In

addition, the determination between acceptable and unacceptable site totals appears to be made on a subjective basis.

Although the histograms indicate that the site totals are not normally distributed, a test for equal site totals using ordinary analysis of variance was performed. For both Lorena's data and the leaf data, we compared two types of sites (Umala vs. Ancoraimes for Lorena's data and Urban vs. Rural for the leaf data). For the soil data, we translated the three-way factorial experimental design into a one-way design and compared the site totals for each of the 24 combinations. We did not perform this test on Neshmi's data because we do not have information regarding the experimental factors so there are no groups to compare. The normal probability diagnostic plots from ANOVA are shown in Figure 3.5. Other than the one small sample, the plot for Lorena's data shows no obvious signs of departure from normality. In contrast, both the soil data and the leaf data show that the assumption of normality is not reasonable.

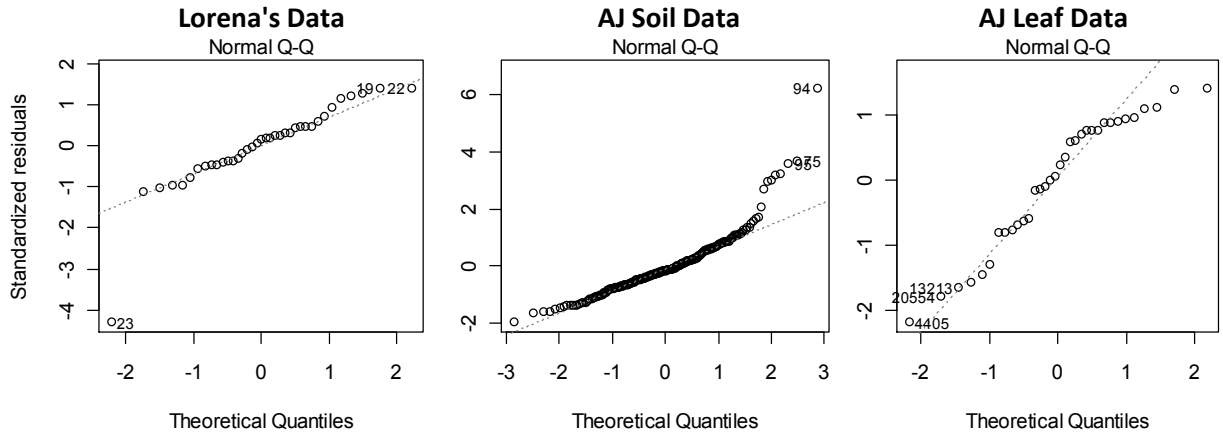


Figure 3.5: Diagnostic Normal Probability Plots from ANOVA

The normal quantile-quantile plots from ordinary analysis of variance indicate that the assumption of normal errors may not hold for some OTU data sets. In particular, both the soil data and the leaf data show obvious departures from normality.

Since the assumption of normality is not satisfied, the ANOVA p-values cannot be trusted for either the soil data or the leaf data. For comparison, we perform the same test using the Kruskal-Wallis procedure. Since this is a nonparametric test, it does not produce a diagnostic normal probability plot. The results are reported in Table 3.2. Note that the leaf data

set is not significant under either test and that Lorena's data set is significant under both tests. The soil data set is only marginally significant under ANOVA (p-value 0.052), but is significant under Kruskal-Wallis (p-value 0.010). Given the appearance of the normal probability plot, a determination based on the nonparametric test may be more valid.

	ANOVA			Kruskal-Wallis		
	df	F	p-value	df	KW	p-value
Lorena's Data	1, 35	4.837	0.035	1	11.440	0.001
Soil Data	23, 214	1.572	0.052	23	41.552	0.010
Leaf Data	1, 32	1.464	0.235	1	0.898	0.344

Table 3.2: Results of Tests for Equal Site Totals

The ANOVA procedure is testing for equality of 'average' site totals under the various experimental conditions, while the Kruskal-Wallis procedure is testing for equality in the complete distributions of site totals. Neither of these procedures is specifically testing for outliers. Delete-1 diagnostic statistics can be computed to gauge the effect of each value, but this will most likely fail to detect clusters of outliers such as the low site totals in the soil data. Further aspects of this approach have not been explored at this time. Section 3.1.5 will present a potential method for adjusting outliers, and Section 3.4 a possible statistical technique for detecting them.

3.1.3. OTU Total Abundances

We now consider the distribution of total counts (abundances) for each OTU across all samples. In the ecological literature, these are called species abundance distributions, and the models can be divided into two categories (Magurran, 2004). Biological models attempt to explain the interdependencies between species and their relation to the environment by relating the types and availability of resources in an environment to the species that inhabit the environment. In contrast, statistical models simply describe the assemblage of species by fitting a probability model to the observed abundance data. Biological models used most often in the ecological literature are based on observed macroscopic organisms, such as the number of birds in a given geographic area, and are applied to pyrosequence data by treating each OTU as a

species. It is currently unknown if the biological mechanisms that control macroscopic species abundances are applicable to microscopic OTU abundances. For this reason, we concentrate on statistical models of species abundance distributions.

Species that are commonly found will necessarily have a different distribution than rare species. For macroscopic organisms, common and rare species are differentiated by their prevalence, that is, the number of samples in which the species is found. Common species (also called core species or residents) occur in many samples and rare species (also called satellites, migrants, or occasional species) are found in fewer samples. Total abundance for common OTUs is usually assumed to follow a lognormal distribution, while rare OTUs are modeled with a log series distribution (Fisher *et al,* 1943). Both of these distributions predict the proportion of OTUs that have a specified total abundance; they are not currently used to predict the abundance of a single OTU. Details of the log series distribution are provided in Appendix B.

Following the method of Magurran and Henderson (2003), core OTUs were separated from satellite OTUs based on a 50% persistence threshold. OTUs that occur in at least 50% of the samples are categorized as core OTUs and the rest are considered rare. This can be visualized in Persistence-Abundance plots, shown in Figure 3.6. For these plots, persistence is defined to be the number of sites in which the OTU occurs and abundance is the maximum abundance of the OTU in any one site.

These plots clearly reflect the differences in the dimensions of these data sets. The leaf data and Lorena's data are roughly equivalent in terms of both number of sites and number of OTUs, but the number of OTUs in Neshmi's data is an order of magnitude larger, and the number of sites in the soil data is roughly six times larger than the other data sets. The dashed vertical lines indicate the 50% persistence threshold, but it is difficult to identify any separation along the horizontal axis, so the classification of rare and core species is ambiguous. This is particularly true for Neshmi's data, in which no separation is apparent. Also note the appearance of the soil data graph, which is substantially different than the other three. This may be caused by the larger number of sites in this data set, but the wide variation in site totals may also be a factor. The 50% persistence threshold appears misplaced in the soil data graph, but there are 238 sites in this data set and the most persistent OTU occurs in only 183 sites. In each of the other data sets, the most persistent OTU occurs in every site.

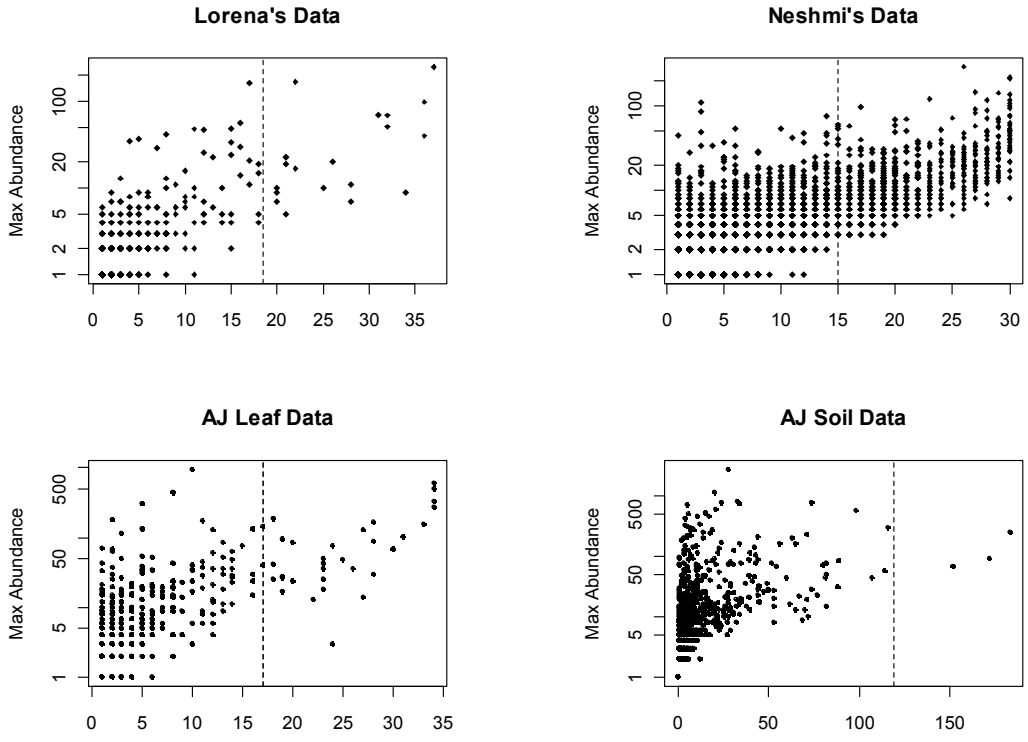


Figure 3.6: Persistence-Abundance Plots

Each point represents one OTU, a surrogate for species. The x-axis is the number of sites in which the OTU is present and the y-axis is the maximum abundance for the OTU in any one site. Note the logarithmic scale on the y-axis, which assists in visually discriminating between the many low abundances and the few large abundances.

Using the 50% threshold, the OTUs in each data set were categorized as either core or rare. To assess the validity of these assignments, a log series distribution is fit to each group and the goodness of fit is evaluated via both a chi-square test and by a Kolmogorov-Smirnov test.

Since abundance data are strongly right-skewed, it is customary in ecological applications to apply the chi-square test to abundance *octaves* rather than to the raw abundance values. This procedure is described in Magurran (2004, page 216) and Krebs (1999, page 430). To create the octaves, the abundance values are log-transformed (using base 2) and the octave boundaries are positive integers. The observed abundances for the OTUs and the abundances predicted by the log series distribution are each binned into to the octaves. To assess the fit, we treat each octave as a category (*i.e.* a table cell) and compare the observed and expected number of OTUs in each category. These calculations were conducted in the R programming language (R Development

Core Team, 2009) using the package *vegan* (Dixon, 2003). The OTU abundance data are very strongly right-skewed, so that some abundance octaves contained no OTUs. Thus some cells contain a value of 0. This presents difficulties for the chi-square test because the asymptotic chi-square distribution is less accurate when cell values are small. To improve the accuracy, a continuity correction of 0.5 was added to each cell before performing the chi-square test.

The Kolmogorov-Smirnov test uses the raw abundance values and does not bin these into octaves. The test statistic is the maximum difference between the empirical cumulative distribution function (ECDF) and the CDF of the proposed log series distribution. This test was conducted in the R Programming Language, using the function `ks.test`. This function utilizes a permutation procedure to obtain a p-value for the test, and the accuracy of the p-value depends on having no tied values (OTU totals) in the data. Among the rare OTUs, which occur infrequently and in small numbers, there are usually a large number of tied OTU totals. Thus the performance characteristics of the Kolmogorov-Smirnov test may need to be scrutinized, especially for rare OTUs.

The p-values for these tests are provided in Table 3.3. Each of these tests is a goodness of fit test to the log series distribution, so that small p-values indicate the log series distribution is not an accurate probability model for the data. We would expect small p-values for the core OTUs and large p-values for the rare OTUs, but this did not occur. Only three p-values are *not* significant at $\alpha = 0.05$, and all three occur for the core OTUs, where we had expected to find small p-values.

The results of the chi-square test suggest that the core OTUs in Lorena's data and AJ Soil data follow a log series distribution, and that all of the other groups of OTUs follow a different distribution. However, it is known that the chi-square test does not perform well when the cell counts are small, and the two largest chi-square p-values (0.9850 and 0.4522) are based on only 4 and 11 OTUs, respectively. When these are dispersed among the abundance classes (octaves), the cell counts are too small for the chi-square test to be reliable. Another difficulty with the chi-square test is that the counts can be too large. For example, there are 20,963 rare OTUs in Neshmi's data, classified into 8 octaves, while Lorena's data has 788 rare OTUs classified into 10 octaves. Neshmi's chi-square test statistic will be much larger than Lorena's, but it will also have fewer degrees of freedom. The result is a p-value of nearly zero, such as those in Table 3.3.

		Core OTUs	Rare OTUs	Combined
Lorena's Data	number of OTUs	11	788	799
	chi-square	0.4522	0	0
	K-S	0.0473	0	0
Neshmi's Data	number of OTUs	657	20,963	21,620
	chi-square	0	0	0
	K-S	0	0	0
AJ Leaf Data	number of OTUs	33	565	598
	chi-square	0	0	0
	K-S	0.0001	0.0001	0
AJ Soil Data	number of OTUs	4	2,418	2,422
	chi-square	0.9850	0	0
	K-S	0.0713	0	0

Table 3.3: P-values for Goodness-of-Fit to Log Series Distribution

Each OTU is classified as either rare or core using a 50% persistence threshold. Each group is tested for a log series distribution using both the chi-square and Kolmogorov-Smirnov tests. The number of OTUs in each group is the sample size for the test. The large number of rare OTUs gives these tests considerable power, resulting in near-zero p-values for all data sets. The opposite is true for core OTUs: the relatively small number of core OTUs may be generating p-values that are too large.

To investigate why the Kolmogorov-Smirnov p-values are so small, we take a closer look at the leaf data. The empirical cumulative distribution functions, shown in Figure 3.7, compare the observed OTU abundances to the abundances predicted from a log series distribution. The maximum difference between the observed and hypothesized CDFs is the test statistic for the Kolmogorov-Smirnov test. While there are clear differences between the ECDFs of the core OTUs, differences in the rare OTUs are obscured due to the large number of rare OTUs, as determined by the 50% persistence threshold. In spite of this, the p-values for these two tests are nearly identical. For the core OTUs, the K-S test statistic is D = 0.4848, with p-value 0.000855 and for the rare OTUs, the K-S test statistic is D = 0.1168, with p-value 0.000897. This p-value, however, relies on having no ties in the data. This is not true for OTU abundance data, particularly for the rare OTUs whose abundance values are usually smaller. For the leaf data, there are 565 rare OTUs, but only 90 distinct abundance values. Thus there are numerous ties in the data, and the accuracy of Kolmogorov-Smirnov test is questionable. Most of the tied values are small abundances (less than 15), so it is possible the removing OTU singletons and additional data pre-processing may reduce the number of ties.

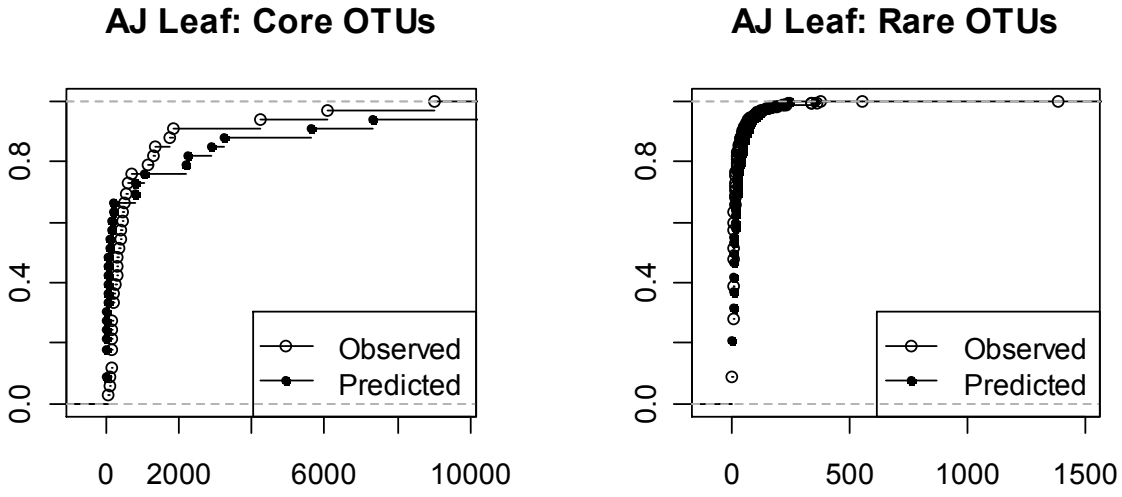


Figure 3.7: Empirical CDFs for OTU Abundance

Predicted values are from the log series distribution. The maximum difference between observed and predicted is the test statistic for the Kolmogorov-Smirnov test. The large number of rare OTUs (565 in the leaf data) visually obscure the differences and give the Kolmogorov-Smirnov test extremely large power for detecting very small differences.

Another difficulty with the Kolmogorov-Smirnov test stems from the computing capabilities of the function `ks.test` in the R Programming Language. For one-sample problems such as ours, `ks.test` requires the vector of observed values and a named probability distribution, along with its associated parameters. The log series distribution is not a standard distribution in the R language, so this test was implemented as a two-sample problem in which the named probability distribution is replaced by a randomly generated vector from the log series distribution. Parameters for the log series distribution were generated from the observed data. We realize that this is not a ideal solution, and the validity of this approach was tested on simulated data.

To explore the accuracy of both the chi-square and Kolmogorov-Smirnov tests, we performed a small simulation. Abundance data for rare OTUs were generated as log series and core OTUs were generated as lognormal. Each of these was tested against a log series distribution using both the chi-square and Kolmogorov-Smirnov tests. Parameters for the lognormal distribution were chosen to match the mean and variance of log series distribution with parameter $\theta = 0.995$. Each test was repeated 100 times, using a randomly generated vector

of length 500 following the designated distribution. The results, given in Table 3.4, are testing the null distribution is the log-series distribution. Both the chi-square and Kolmogorov-Smirnov tests performed well, but the chi-square test has significance level 0.08 instead of 0.05. Not surprisingly, the chi-square test also has higher power for detecting the lognormal distribution. Based on these limited results, we conclude that both of these tests are performing adequately, provided there are no ties in the data. For Type I error control we prefer the Kolmogorov-Smirnov test.

	proportion of rejected tests	
True Distribution	χ^2	K-S
log series (θ=0.995)	8%	0%
lognormal (μ=40, σ=80)	96%	83%

Table 3.4: Goodness of Fit Tests on Simulated Data

Both the chi-square and Kolmogorov-Smirnov tests have sufficient power for detecting that a randomly generated lognormal sample is not a log series, but the chi-square test is level 0.08 instead of the intended 0.05.

Since both tests seem to be adequate when applied to simulated data, we return to the four OTU data sets and visually assess the fit by comparing the observed and fitted values. These graphs, shown in Figure 3.8, indicate that the log series distribution fits the rare OTUs better than the core OTUs. However, the log series distribution fails to capture the extreme height in the lower octaves of the rare OTUs. This lack of fit may diminish when the singletons are removed from the data. Although not conclusive, these results bring into question the validity of using the log series distribution to model rare OTU abundances. It is possible that a different distribution or a mixture distribution is needed, or perhaps the size of the octaves could be adjusted (by altering the base of the logarithm) to obtain a better fit. A similar procedure could be used to assess the suitability of the lognormal distribution for core OTUs, but this issue has not been addressed at this time.

These tests highlight some critical differences between OTU data sets and the data collected on macroscopic species, so that methods applied to macroscopic data may not be applicable to OTU data. Some key differences are

- Customary probability models for species abundance distributions do not capture the extreme variability in OTU data.
- The 50% persistence threshold for common/rare classification does not accommodate the large proportion of zeros in OTU data.
- Traditional goodness-of-fit test are inadequate due to the high dimensionality and sparseness of OTU data.

These results illustrate that the procedures and assumptions regarding macroscopic species abundance data cannot be directly applied, without consideration, to OTU abundance data. Perhaps the 50% threshold needs to be shifted to better discriminate between common and rare OTUs, or perhaps a different criterion is needed. This is discussed in more detail in Section 3.3. Another concern is the suitability of the chi-square and Kolmogorov-Smirnov tests to perform accurately on OTU data, which is characterized by its extreme skewness and large dimension. The difficulties in evaluating the goodness of fit tests have been addressed by other authors, for example Mouillot and Lepretre (2000), who note that the observed data often fit either all of the proposed probability models or none at all.

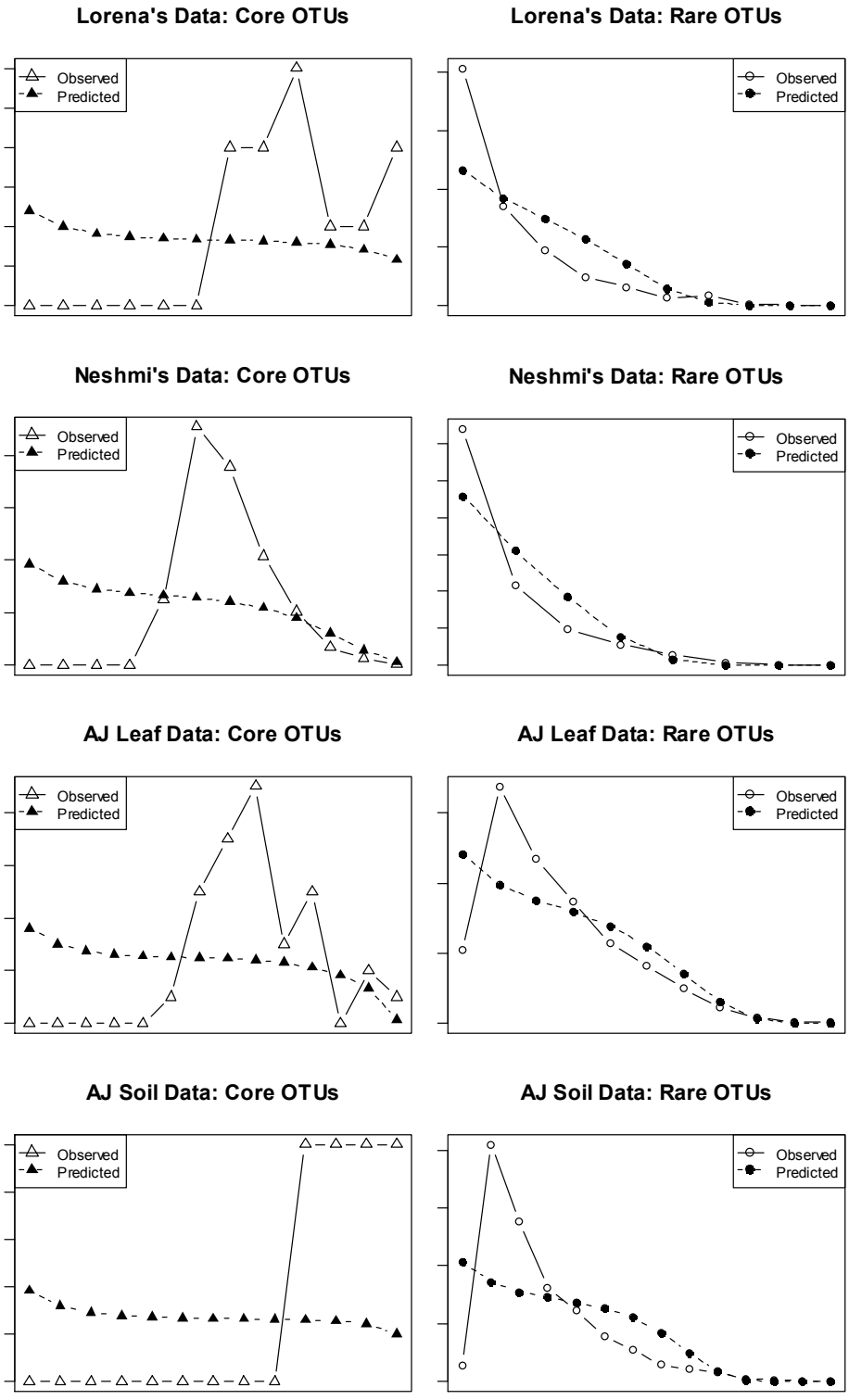


Figure 3.8: Observed vs. Predicted OTU Abundance Patterns

Abundance classes (octaves) are represented on the x-axis with number of species on the y-axis. Predicted values are from a log series distribution.

3.1.4. OTU Singletons

Singletons occur when a particular DNA sequence is detected only once at a site. As discussed in Section 2.2, the presence of singletons in a data set can have a profound impact on the estimate of species richness and subsequent measures of diversity, and thus warrant special attention. They are generally perceived to be the result of sequencing errors, but they could represent an extremely rare species. Typically, OTU data sets contain only a small proportion of nonzero counts, but many of these are singletons. In Lorena's data, for example, only 8.7% of the counts are greater than zero, but over half of these (56.2%) are singletons.

To explore the impact of singletons, we examine the soil data. This data set was chosen because it contains the smallest percentage of nonzero counts (3.7%), and nearly half of these are singletons. For each singleton, we examine the context of the singleton by considering the counts for the same OTU at the other sites. We are concerned with the number of other sites at which this OTU occurs and whether or not these other occurrences are also singletons. This information can serve as a basis for deciding whether the OTU represents a sequencing error (which should be removed from the data) or a 'real' OTU which should be kept in the data.

The soil data contains a total of 10,442 singletons. Each of the 238 sites has at least one singleton, and 2001 of the 2422 OTUs occur as a singleton in at least one site. There are 657 OTUs that occur *only* as singletons, that is, these OTUs were either not detected or detected only once at each site. There are an additional 1344 OTUs that occur in some sites as singletons, but also occur in greater number at some sites. The remaining 421 OTUs never occur as singletons.

We first examine the 657 OTUs that occur only as singletons. The combined count for these OTUs is 1,833 of the total 10,442 singleton counts. The distribution of occurrences of these 657 OTUs is shown in Figure 3.9. Note that 56 of these OTUs occur at only one site. Thus in the entire data set, each of these OTUs occur only once and they occur as a singleton. A strong argument could be made for attributing these to sequencing errors, and eliminate these OTUs from the data set. It is less clear, however, how to interpret remaining OTUs that occur at two or more sites, but each occurrence is a singleton. For example, there is one OTU that occurs (with a count of one) at 12 sites. It is unclear whether this should be interpreted as an extremely rare OTU (that is present at 12 sites) or as a sequencing error that occurred 12 times.

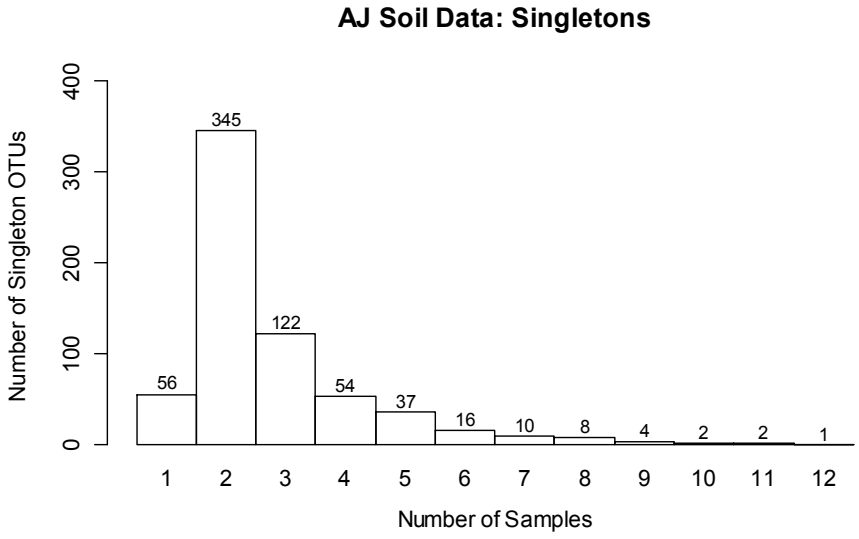


Figure 3.9: OTU Singletons in the Soil Data

Distribution of the 657 OTUs that occur only as singletons in the soil data. For example, there are 56 singleton OTUs that each occurred exactly one time in exactly one sample (one site), while 345 OTUs occurred exactly one time in two different sites. Note that one OTU was detected at exactly 12 sites and was recorded as a singleton at all 12 sites.

A more complicated pattern of singletons emerges when we broaden our attention to the 1344 OTUs that occur as singletons, but also occur at other sites in greater numbers. These singletons account for 8,609 of the total 10,442 singletons. Since these OTUs occur in other sites with larger counts, it is of interest to compare the occurrence of singletons and nonsingletons for these OTUs. These are shown in Table 3.5. This characterization is harder to describe, but in general, OTUs that occur as a singleton in a small number of samples are also likely to occur as a nonsingleton in a small number of samples. For example, the upper left cell in Table 3.5 indicates that 752 OTUs occur as singletons in 5 or fewer sites, and these same OTUs occur as nonsingletons in 5 or fewer sites. The bottom right cell indicates that 51 OTUs occur as singletons in more than 20 sites, and also occur as nonsingletons in more than 20 sites.

		Number of Sites at which OTU occurs with count >1					
		1-5	6-10	11-15	15-20	>20	Total
Number of Sites at which OTU occurs as a Singleton	1-5	752	61	11	9	4	837
	6-10	149	59	23	9	17	257
	11-15	31	38	20	6	31	126
	16-20	5	10	7	8	24	54
	>20	2	7	6	4	51	70
	Total	939	175	67	36	127	1344

Table 3.5: Number of Nonzero Counts for Mixed OTUs in the Soil Data

In the soil data, 1,344 of the 2,422 OTUs occur as a singleton in at least one site and also occur as a nonsingleton in at least one site. These are called mixed OTUs, and are classified according to the number of times (number of sites) they occur as a singleton and the number of times they occur as a nonsingleton. OTUs that occur often as a singleton but rarely as a nonsingleton may represent sequencing errors, which should be removed from the data.

It is unclear what guidelines should be used to determine whether these singletons should be kept or discarded. It seems intuitive that OTUs that occur in a small number of sites, and occur only as singletons at those sites, are most likely the result of sequencing errors and should be removed from the data. But if an OTU occurs in some sites as a singleton and in other sites as a nonsingleton, the decision is not as straightforward. The options include:

(1) Remove all individual counts of one and then remove the OTUs that are no longer observed at any site. For the soil data, this would eliminate 657 of the 2,422 OTUs and reduce the total count in the data set by 10,442, a reduction of 6.8%.

(2) Remove all OTUs that occur only as singletons. For the soil data, this would eliminate 657 OTUs, but would remove the counts only for these OTUs. The reduction in total count is 1833, or 1.2%.

(3) Remove all OTUs that occur only as singletons and meet another minimum threshold requirement. The additional requirement could specify a the number of sites at which the OTU occurs, or could specify a minimum total count for the OTU.

These options are further complicated by the fact that a sequencing error could be duplicated, resulting in counts of two or more for a 'phantom' OTU. This could occur, for example, if the target DNA (the 'true' OTU) contains homopolymers which are routinely mis-called by the base calling software, as discussed in Section 1.2.2.2. Multiple errors could

also be the result of variation in PCR amplification. With these considerations, it is not only the singletons that are problematic, but other small counts may need to be scrutinized as well.

3.1.5. Outliers in Individual Abundances

Much of the preceding work has focused on marginal analysis, that is, row (site) totals and column (OTU) totals of the abundance matrix. The individual entries in the matrix, representing the abundance for a particular OTU at a particular site, are more difficult to characterize because the matrix contains so many zeros. The nonzero values are highly skewed, with numerous small values punctuated by occasional extremely large values. A comparison of the 95th to 100th percentiles, Table 3.6, clarifies this point. For example, in the soil data there are 21,465 nonzero values in the matrix. Among the nonzeros, 95% of the values are 19 or less, and 99% are less than 105, but there is at least one value of 2739.

Percentile	Lorena	Neshmi	AJ Leaf	AJ Soil
95%	29	9	78	19
96%	37	10	97	25
97%	53	13	124	34
98%	81	17	175	54
99%	122	26	296	105
maximum	253	292	962	2739

Table 3.6: Upper Percentiles of Individual Nonzero Abundances

The extreme skewness of individual OTU abundances can be seen in the upper percentiles of the distribution. In Neshmi's data, for example, 95% of the nonzero entries in the OTU abundance matrix are less than 9, but the largest value is 292. Similar disparities occur in all four data sets.

A plot of the 50 largest individual values in the soil data, shown in Figure 3.10, clearly shows the separation between the largest and second-largest value in this data set. It is difficult to interpret such a large value, when the majority of other values are so small. Could this represent an OTU that has completely dominated a site, or is this value artificially inflated as a result of, perhaps, PCR amplification? Regardless of the cause, the existence of a small number of extremely large values has implications in all statistical analyses. The values can cause excessive variation in both the row (site) totals and column (OTU) totals, which creates difficulty

in fitting probability models to these totals. Excessively large values can also distort distances between the sites or between the OTUs.

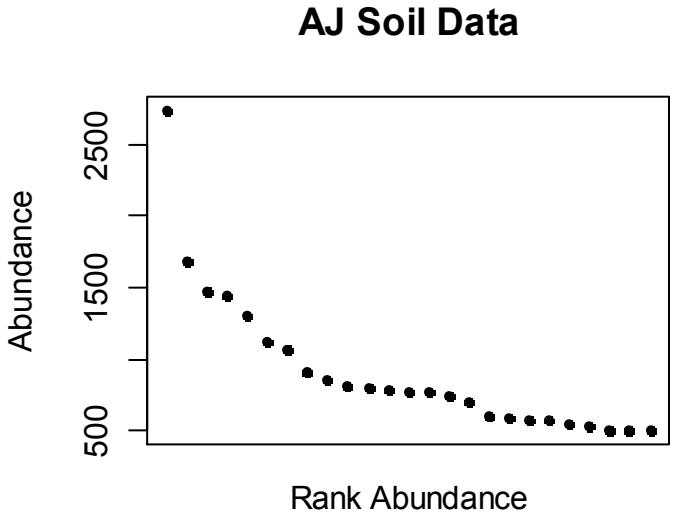


Figure 3.10: Twenty-five Largest Individual Abundances in the Soil Data
Each point represents an individual OTU abundance (the count for a particular OTU at a particular site). These 25 abundance values comprise only 0.1% of all the nonzero values in the soil data. Excessively large values will dominate the statistical analysis and obscure the contribution of the lesser abundance values to ecological diversity measures.

One possible method for mitigating the effect of excessively large counts is to simply use rank abundance instead of the actual abundance values. This approach is not deemed practical for OTU data, given the large number of ties among the lower abundance values. As an alternative approach, we develop an algorithm that reduces the values of the excessive counts while maintaining the rank order relationship among all related counts. For our purposes, a 'related' count is any count for (1) the same OTU or (2) the same site or (3) any site that shares at least one environmental and/or experimental condition with the large count. For example, if one large count was detected for OTU *a* at a site that had been subjected to treatment conditions T_1 and T_2, the related counts include all counts for OTU *a* (regardless of the site), and all counts for all OTUs at the sites that experienced either treatment T_1 or T_2. For brevity, the collection of related counts is called a 'cohort'.

Algorithm for Reducing Overly Large Individual Counts

1. Specify the user inputs:
   a) N = the number of large counts to examine
   b) P = the proportion of separation to maintain between counts
   c) the factors that define the cohorts
2. Identify the N largest individual OTU counts.
3. For each large count C,
   a) Identify the next largest count in the cohort; call this count M
   b) If the ratio C/M is less than 1 + P, then do not change the current count. Go to Step 3 and get the next "big" count.
   c) If the ratio C/M is greater than 1 + P, then change the current count so that it is P% more than M .
4. Go to Step 3 and get the next "large" count.
5. When all N counts have been processed, repeat Step 3. Continue repeating Step 3 until there are no changes to any of the N counts.

This algorithm was applied to the soil data. The twelve largest counts in the data set (which are the counts greater than 1000) were targeted for potential reduction, maintaining a separation of P = 10%. After six iterations, the seven largest counts were reduced, but the remaining counts were unchanged. A summary of each iteration is given in Table 3.7. Six of the changed counts were for OTU 2023, and the total count for this OTU was reduced by 1925. The only other OTU affected is OTU 1173, which was reduced by 114. Each of the seven changed values occurred in different sites, so only seven site totals were affected. Changes to the site totals, shown in Figure 3.11, indicate a clear reduction in the skewness of this distribution. The reduction may be insufficient, however, in that the lesser counts may still be dwarfed by the excessive counts. A more robust reduction technique is presented in Section 4.1.

			Iterations						
Site	OTU	Count	1	2	3	4	5	6	Reduction
06_4_040	2023	2739	**1849**	**1781**	1781	1781	**1708**	**1621**	1118
05_3_040	2023	1681	**1619**	1619	1619	**1553**	**1474**	1474	207
09_2_060	2023	1472	1472	1472	**1412**	**1340**	1340	1340	132
06_4_060	2023	1437	**1431**	**1284**	**1218**	1218	1218	1218	219
06_1_040	2023	1301	**1167**	**1107**	1107	1107	1107	1107	194
02_3_020	1173	1121	**1007**	1007	1007	1007	1007	1007	114
10_4_040	2023	1061	**1006**	1006	1006	1006	1006	1006	55
09_2_020	2023	915	915	915	915	915	915	915	0
09_2_040	2023	861	861	861	861	861	861	861	0
02_4_040	2269	808	808	808	808	808	808	808	0
09_4_060	2023	797	797	797	797	797	797	797	0
07_3_100	2072	779	779	779	779	779	779	779	0

Table 3.7: Results of Algorithm for Reducing Large Counts

After 6 iterations, the largest individual count in the soil data was reduced from 2739 to 1621, and the second-largest count was reduced from 1681 to 1474. Twelve individual counts were targeted for reduction, but only seven were actually reduced.

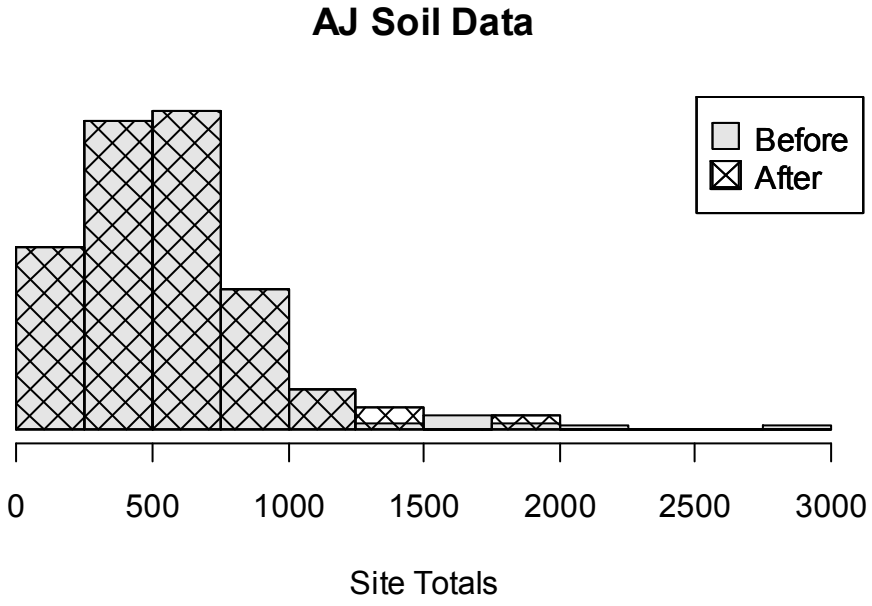


Figure 3.11: Downward Shift of Site Total Abundances

The shaded histogram represents the distribution of total site abundances before the adjustment and the histogram with the cross-hatches indicates the distribution after the adjustment.

3.2. Applying Logratio Analysis to OTU Abundance Data

As discussed in Section 3.1.1, the mechanisms by which OTU abundance are generated should produce site totals that are roughly equivalent. To accommodate random fluctuations in these totals, some researchers prefer to analyze relative frequencies rather than the raw abundance values. The relative frequencies are defined to be the proportional abundances at each site, so that the total relative abundance at each site is equal to one. Whenever relative frequencies are used, the analysis should take into account the constrained nature of these values. This is accomplished with compositional data analysis, in particular logratio analysis, as presented in Appendix D. There are many considerations in applying logratio analysis to OTU data. Some of these issues are addressed below.

3.2.1. Zeros

Logratio analysis requires that all data values be strictly positive, but OTU data typically contain 90% or more zeros. Logically, it seems that we could combine (amalgamate) some of the OTUs to reduce the number of zeros. It also seems logical that we could use taxonomic classification (species-genus-family or higher orders) to perform this amalgamation. Unfortunately, it seems that there is not currently a clear mapping between DNA sequences and taxonomy and that amalgamation would introduce too much uncertainty in the data (Jumpponen, personal communication, Fall 2011). Another method for eliminating zeros is simply to replace each zero with a small positive value. This approach treats each zero as a missing value, and imputes an appropriate replacement value based on the remaining data. Although many zero-replacement strategies have been proposed, they address only 'rounded' zeros. That is, they assume the actual abundance for each OTU at each site is a positive value, but is too small to be detected. Many of the zeros in an OTU data set are 'structural' zeros, that is, these a zero occurs because an OTU is not present at a site. Thus a zero in an OTU data set conveys important information, and should not routinely be replaced with a positive value.

3.2.2. Spurious Correlation

The primary reason for considering logratio analysis for relative frequency data is that the data are constrained by a common sum (the site total), so that the relative frequencies are not independent even when the original abundances are independent. The dependency created by the

sum-to-one constraint is called spurious correlation, and the methods of logratio analysis are designed to accommodate this correlation so that any 'true' correlation can be revealed.

The effect of spurious correlation can be seen in the following example. Let the random variables X, Y and Z represent the raw abundances for three OTUs and suppose that they are mutually independent. Divide each raw abundance by the sum (X+Y+Z), to create the compositional vector of relative frequencies (C1, C2, C3). Any correlation between the C's will be spurious correlation, since the elements of the original vector are presumed to be independent. To illustrate, we simulate 1000 independent and identically distributed samples of size 200 for each of X, Y and Z, following a gamma distribution with shape 2 and scale 250. (These parameters were chosen to mimic the shape of OTU abundance data). Each (X, Y, Z) triple represents one of the 200 sites, but when viewed across sites the collection of 200 values for X represent a random sample of 200 abundances for the first OTU, with a similar interpretation for Y and Z. We use these 200 values to calculate the pairwise correlations between the original random variables and the pairwise correlations between the corresponding relative frequencies. This provides 1,000 simulated values for each correlation. The distributions of these correlations are shown in Figure 3.12. Since the raw abundances (X, Y, Z) are independent, their pairwise correlations are, as expected, centered at 0. In contrast, the correlation between relative abundances are all negative, and centered at approximately -0.5. This is the spurious correlation.

The most extreme case of spurious correlation occurs when there are only two OTUs. In this case, the vector of raw abundances at a site is (x_1, x_2) and the relative abundances are $(p_1, p_2) = \left(\frac{x_1}{x_1+x_2}, \frac{x_2}{x_1+x_2}\right)$. Regardless of the relationship between x_1 and x_2, we have $p_1 + p_2 = 1$ so that $\text{cor}(p_1, p_2) = -1$. The preceding simulation illustrates that when a third OTU is included, the correlation between proportions is not as strong, but is still negative. We know that typical OTU data sets contains hundreds, perhaps thousands, of OTUs. Does this large dimension reduce spurious correlation to a negligible value? We explore this question with another simulation.

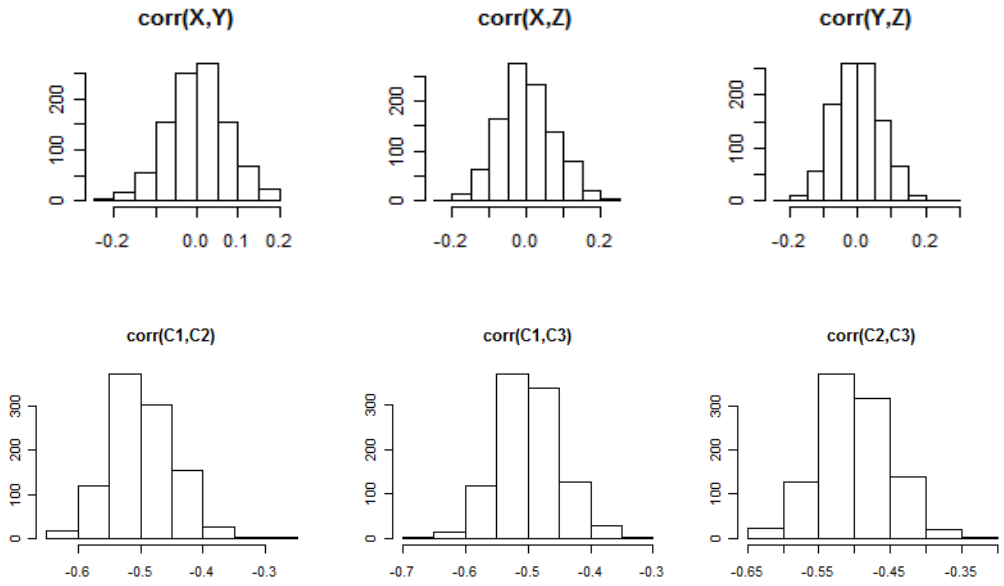


Figure 3.12: Spurious Correlation for Three OTUs

The top row of histograms show the sampling distribution of the correlation between independent vectors, based on 1000 replications of three random vectors (X, Y, Z). Each vector has length 200. The bottom row of histograms show the sampling distribution of the spurious correlation between the elements of the corresponding compositional data vectors.

To investigate the degree to which the number of OTUs affects the magnitude of spurious correlation, we performed a series of simulations. The number of OTUs was varied from 50 to 2000. Each simulation contained 200 sites and each OTU abundance at each site was generated independently from a mixture distribution

$$\pi_0 \cdot gamma(0.1,\ 1) + (1-\pi_0) \cdot gamma(0.1, 50)$$

with $\pi_0 = 0.80$. These parameters were chosen to mimic the leaf data, discussed in Section 3.1. For each collection of OTUs, the correlation between raw abundances is compared to the correlation between the corresponding proportions. The results, shown in Figure 3.13, are for Pearson's correlation. Similar results were obtained for Spearman's correlation and for Kendall's tau. These graphs illustrate that, as the number of OTUs increases, the correlation between proportions approaches the raw correlations. Thus the spurious correlation diminishes as the number of OTUs increases.

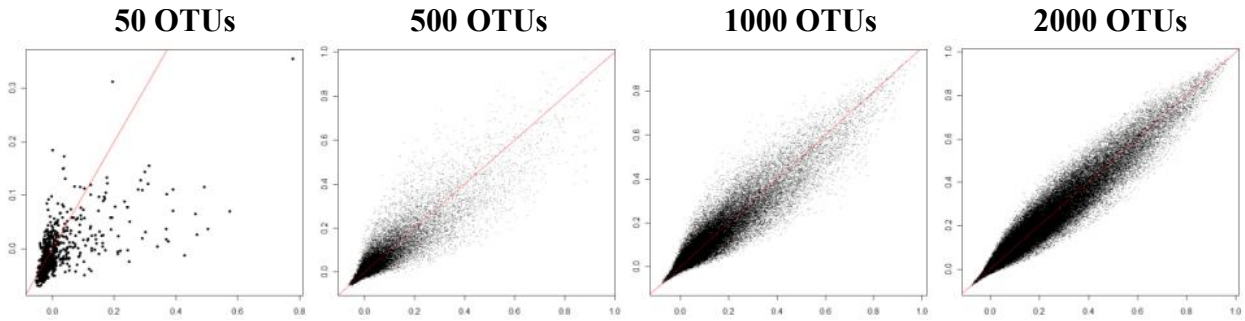


Figure 3.13: Simulated Spurious Correlation as the Number of OTUs Increases
Each point represents one pair of OTUs. Correlation between raw abundances is on the x axis and the correlation between the proportions (relative frequencies) is on the y axis. The line indicates where these correlations are equal.

In addition to the simulation results, under certain distributional assumptions it is possible to express the spurious correlation as a closed form expression.

Result 3.2:

Assume $X_j \sim gamma(\alpha_j, \beta)$, $j = 1, 2, \ldots, S$, with X_j and $X_{j'}$ independent for $j \neq j'$ and define $P_j = \dfrac{X_j}{\sum_{i=1}^{S} X_i}$.

Then $\text{corr}\left(P_i, P_j\right) = \dfrac{-\sqrt{\alpha_i \alpha_j}}{\sqrt{\left(\alpha_+ - \alpha_i\right)\left(\alpha_+ - \alpha_j\right)}}$, where $\alpha_+ = \sum_{k=1}^{S} \alpha_k$

The derivation of this result is given in Appendix D. Note that the correlation is negative and that it does not depend on the common scale parameter. Both of these results are intuitive. A negative correlation occurs because the relative frequencies must sum to 1. Also, these proportions are ratios of gamma random variables with the same scale parameter, so their distributions and hence their correlations should not depend on the scale parameter. Also note that each α_i is positive, so that α_+ is strictly increasing in S (the number of OTUs), but it is not guaranteed to increase without bound. For some values of the α_i, this sum may converge to a constant. For any fixed i and j, the numerator is constant while the denominator will increase

with *S*. Thus the absolute value of the correlation will decrease as *S* increases, but it is still unclear under what conditions the correlation will go to 0.

3.2.3. Other Considerations

In addition to the preponderance of zeros and the effects of spurious correlation, there are other challenges with applying logratio analysis to OTU data. One such challenge is the assumption that logratio-transformed relative abundances follow a multivariate normal distribution. Given the extreme skewness of abundance data, it is unlikely that any transformation will induce normality, although this has not been formally examined. Another challenge is a result of the large dimension of OTU data sets, so that some operations are not feasible without extended computing power. For example, one version of a logratio variation matrix requires examining the ratio of every pair of OTUs. In Neshmi's data there are 21,620 OTUs and over 200 million pairs of OTUs. This is beyond the capability of most computers.

In view of all of these difficulties, it has been decided that logratio analysis of OTU data should not be pursued at this time. This decision is based on two principal considerations: (1) spurious correlation, which is one primary reason for performing logratio analysis, is only a minor obstacle in large dimensional OTU data sets; and (2) replacing all the zeros with positive values would introduce too much uncertainty in subsequent analysis.

3.3. Rare vs. Common OTUs

As described in Section 2.5, it is typical to classify OTUs as either common or rare, based on their prevalence in the observed data. This is a logical classification, since the biological and environmental mechanisms that govern common species are distinctly different than the mechanisms that govern rare species. In particular, it has been noted that "biological factors underpinned the relative abundance of the core [common] species, whereas random dispersal was more important in structuring the satellite [rare] species." (van der Gast, *et al.*, 2011, page 781) This suggests that the abundance patterns for core and rare species follow different probability models, although the precise form of appropriate models remains a source of much debate. This is discussed in Section 3.4 The assignment of OTUs into common and rare groups is also important for measuring changes in the health of a habitat, since "[e]cologically relevant shifts in abundances probably occur predominantly among the core [common] members that are by definition well established in the system" (Unterseher *et al*., 2011). Thus an accurate description of OTU abundance data requires a reliable method for identifying rare and core OTUs.

The 50% persistence threshold for macroscopic species, recommended by Magurran and Henderson (2003) and described in Section 2.5, seems ill-suited for OTU data primarily because the persistence patterns for microscopic species are very different from macroscopic species. In macroscopic data sets, it is customary to have many species that occur in every sample, but this is not true for microscopic data. In general, the large dimensions of OTU data sets and the excessive number of zeros distort the persistence threshold. An example of this occurs in the soil data in which only four OTUs are present in at least 50% of the sites (samples). These four OTUs would be classified as common, and the remaining 2,418 OTUs would be classified as rare. Shifting the persistence threshold to 40% generates only 13 common OTUs, which is only approximately $\frac{1}{2}\%$ of the total number of OTUs.

When using Magurran and Henderson's persistence threshold criterion, only the persistence is used to classify an OTU. The abundance values for the OTU are ignored. It seems reasonable that an OTU classification criterion should use both the persistence and abundance, and that OTUs with low persistence and irregular abundance patterns should be classified as rare, while those OTUs with high persistence and a more even distribution of abundances should be

classified as common. We propose a new method, based on the Gini index, which is described in the following section.

3.3.1. Lorenz Curve

The Lorenz curve is used in financial applications to model the distribution of wealth (or income) in a population. It is designed to accommodate distributions with long right tails, and is therefore particularly well suited for OTU abundance data. A Lorenz curve describes the inequity, or unevenness, in a distribution and we believe it may prove useful to differentiate between rare and common OTUs.

A Lorenz curve is defined by two parametric functions of the probability density of a continuous random variable W. In financial models, W is typically the wealth of individuals (or households) in a population. To apply this to OTU data, we let W represent the abundances for one OTU across several sites. Suppose W has pdf *f* and cdf *F*, with support (0, ∞) and mean μ. Each value of *w* in the support generates one (*x, y*) pair on the Lorenz curve, defined by

$$x(w) = F(w) = \int_0^w f(t)\,dt$$

$$y(w) = \frac{1}{\mu}\int_0^w t \cdot f(t)\,dt$$

Note that both *x* and *y* are bounded in the interval [0, 1]. These values are sometimes converted to percentages. In a financial situation, the point (0.25, 0.1) indicates that the 25% of the population with the lowest wealth controls 10% of the total wealth. When this is applied to OTU data (in particular, *one* OTU across several sites), this point indicates that the smallest 25% of the sites contain 10% of the total abundance for this OTU. In this sense, a 'small' site is one that has a low abundance for this OTU.

If the abundances are perfectly evenly distributed, then the lowest 10% of the sites will contain 10% of the total abundance, the lowest 25% of the sites will contain 25% of the total abundance, and so on. Thus the line $y = x$ is called the line of perfect equality. Any inequity in the distribution will cause the Lorenz curve to fall below this line, as indicated in Figure 3.14. Greater inequity creates greater distance between the curve and the line of perfect equality.

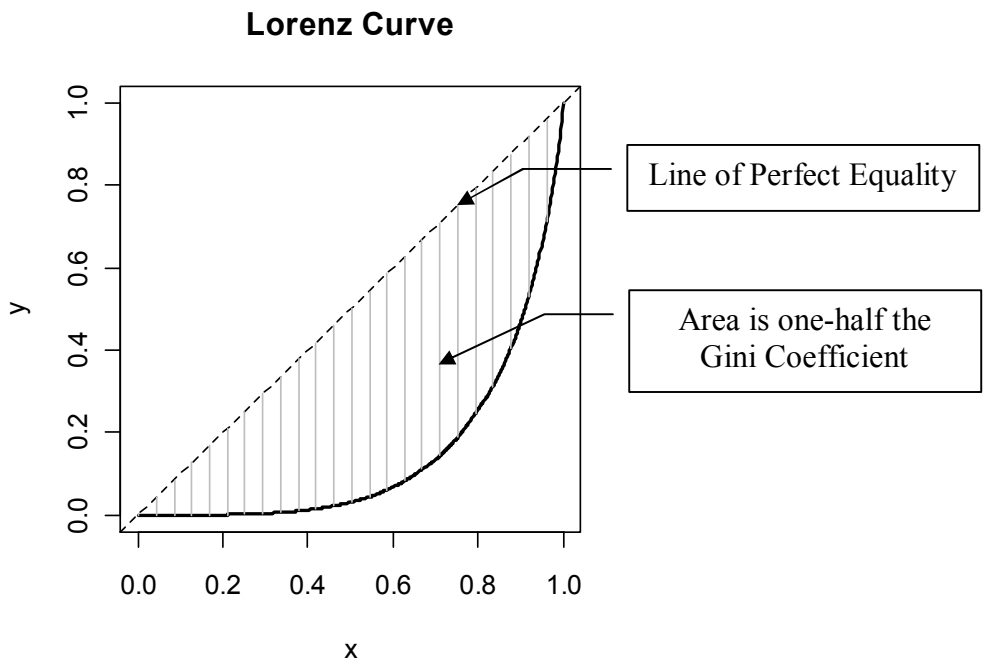


Figure 3.14: A Lorenz Curve

The curve represents one OTU across several sites. The x axis represents the accumulated proportion of sites and the y-axis represents the accumulated proportion of abundance at those sites. For example, the point (0.8, 0.2) indicates that 80% of the sites contain 20% of the total abundance for this OTU.

3.3.2. Gini Index and Asymmetry Coefficient

The area between the Lorenz curve and the line of perfect equality is a measure of the total amount of inequity. Since this area is always between 0 and 0.5, it is customarily doubled so that it is between 0 and 1. The doubled area is called the Gini index (or Gini coefficient). We postulate that the Gini index can be used as a measure of unevenness in the abundances of an OTU and can therefore serve to discriminate rare OTUs from common ones. Rare OTUs occur in few sites and their abundances can be irregular. Common OTUs, in contrast, occur in many sites and their abundances tend to be more stable (more even). Measures of variability based on sample moments, such as the standard deviation and skewness, are affected by the magnitude of the observed abundances. Given the extreme variation in OTU abundance data, comparisons based on sample moments may be unreliable.

The Gini index is but one univariate measure based on the Lorenz curve. Gastwirth *et al.* (2005) describe Gini's mean difference and the Gini index of concentration, but the

applicability of these measures to OTU data have not been explored at this time. Another measure, the Lorenz Asymmetry Coefficient (LAC) proposed by Damgaard and Weiner (2000), may prove useful in distinguishing rare and common OTUs. The LAC is calculated from one point on the Lorenz curve, specifically the point (x_0, y_0) at which the slope of the Lorenz curve equals 1. This point is chosen because, for the subpopulation of sites in a neighborhood of this point, the abundances are distributed equally. (The slope 1 matches the slope of the line of perfect equality.) This point (x_0, y_0) is generated from the parametric equations evaluated at the value μ, the mean of the distribution of abundances. In other words, $x_0 = x(\mu) = \int_0^{\mu} f(t)\,dt$ and $y_0 = y(\mu) = \frac{1}{\mu}\int_0^{\mu} t \cdot f(t)\,dt$. The Lorenz Asymmetry Coefficient is defined to be sum of the two coordinates, $x_0 + y_0$, and the Lorenz curve is said to be symmetric if the LAC equals 1. A geometric interpretation is based on the line of symmetry, which is perpendicular to the line of perfect equality. The equation for the line of symmetry is $x + y = 1$, so if the Lorenz curve is perfectly symmetric, the point (x_0, y_0) will be on this line. The concept of symmetry can be visualized by rotating the graph so that the line of perfect equality is horizontal and the line of symmetry is vertical. If the curve is symmetric, the graph will be mirror images around the line of symmetry. It should be noted that the Lorenz Asymmetry Coefficient does not measure the symmetry in the original distribution of abundances. Instead, it measures the symmetry in the accumulated abundances, as depicted in the Lorenz curve. One very interesting characteristic of the LAC is that, if the underlying distribution of abundances is lognormal, then the theoretical value for the LAC is 1. This fact may be useful for testing whether the abundances follow a lognormal distribution (and the OTU would therefore be designated as a common OTU).

Two simulations may help to clarify these concepts. In the first simulation, abundance data for one OTU and 500 sites were generated from a gamma distribution with shape parameter 0.3 and scale parameter 0.05 (so the mean is 6). The histogram and resulting Lorenz curve are shown in Figure 3.15. For comparison, the graphs shown in Figure 3.16 are based on 500 observations from a gamma distribution with shape 2 and scale 0.01 For both of these data sets, the distribution of abundances is skewed, but the skew is less pronounced in the latter

distribution. This is reflected in the corresponding Lorenz curves, and in particular the Gini coefficient, which is 0.720 for the distribution that is more skewed and 0.353 for the distribution

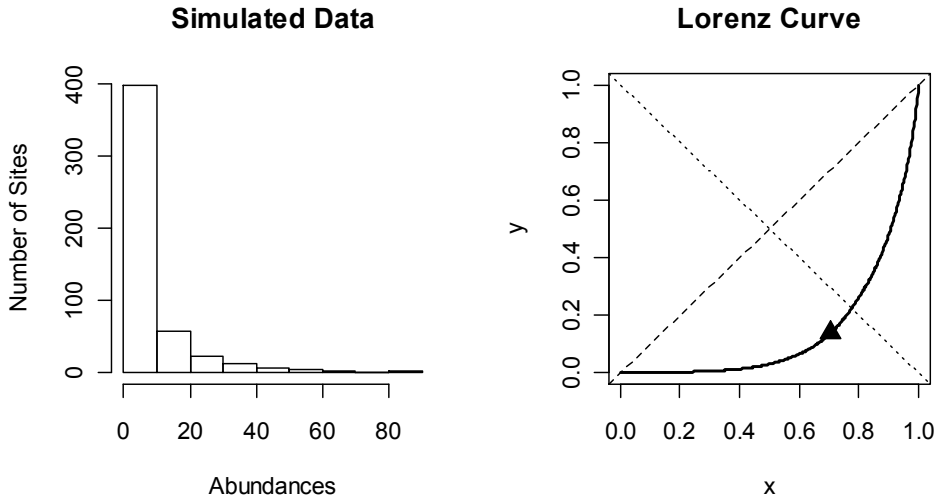


Figure 3.15: Lorenz Curve for Gamma (0.3, 0.05)

The Lorenz curve is generated from the 500 simulated OTU abundance values shown at left. The abundances follow a gamma distribution with shape parameter 0.3 and scale parameter 0.05. The Gini coefficient is 0.720 and the asymmetry coefficient is 0.842

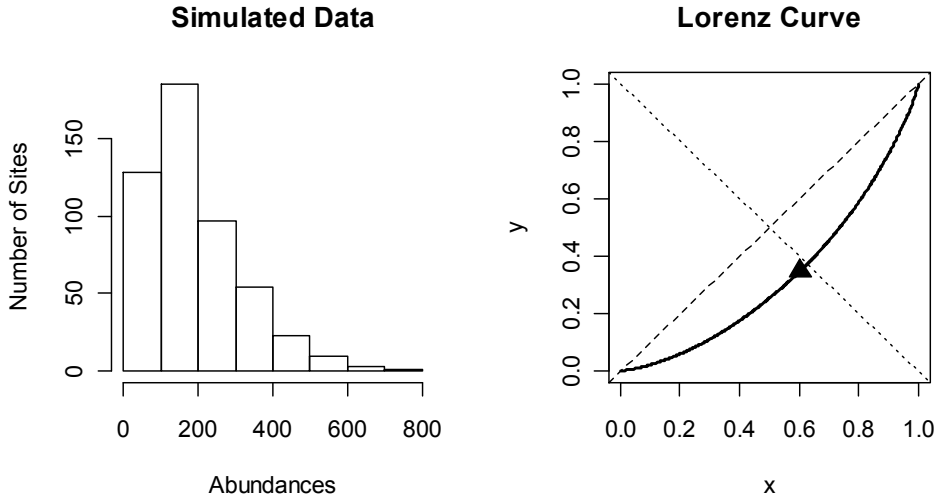


Figure 3.16: Lorenz Curve for Gamma (2, 0.01)

The Lorenz curve is generated from the 500 simulated OTU abundance values shown at left. The abundances follow a gamma distribution with shape parameter 2 and scale parameter 0.01. The Gini coefficient is 0.353 and the asymmetry coefficient is 0.950

that is less skewed. A smaller value for Gini coefficient indicates a more equitable distribution. The asymmetry coefficients, 0.842 and 0.950, respectively, also reflect the differences in skewness of these two distributions, but a more accurate interpretation of these coefficients is that they reflect the differences in *accumulated* abundances. This is shown in Figure 3.17.

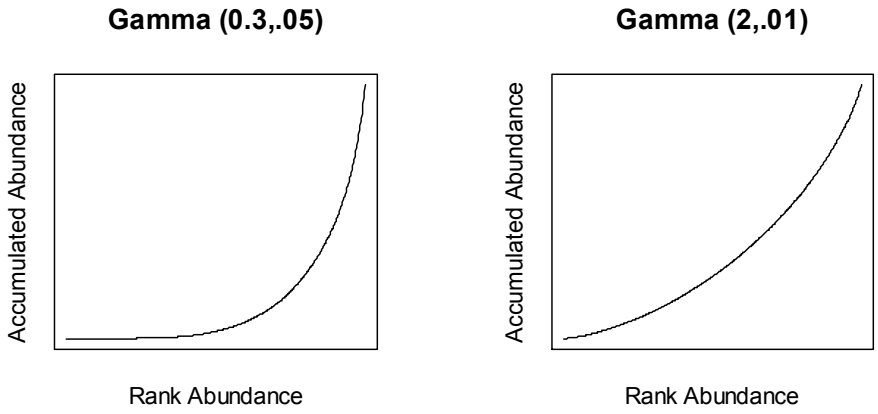


Figure 3.17: Curvature in Rank Abundance Plots

The x axis represents the rank of the abundance values, from smallest to largest. The y axis is the cumulative total abundance. The Lorenz asymmetry coefficient captures the curvature of this graph. The graph on the left has Lorenz asymmetry coefficient equal to 0.842, while the one on the right is 0.950. This coefficient is in the range [0, 2], and values at each extreme are the result of a nearly straight line.

In order to assess whether the Gini coefficient or the Lorenz asymmetry coefficient are suitable statistics for conducting hypothesis tests, we need to evaluate the characteristics of their sampling distributions. The results of several simulations are shown on the following pages. These results are limited in scope, since the intent is to determine if there is a preference for one of these measures for use in distinguishing rare vs. common OTUs. We compare four distributions for the abundances: exponential with parameter 0.01, lognormal with parameters 1.8 and 1.5, Pareto with parameters 1 and 1.00000025, and gamma with parameters 2 and 0.1. The parameters were chosen arbitrarily to generate highly skewed distributions typical of OTU abundance data. For each of these distributions, we considered two sample sizes: 30 and 150. For each of the eight combinations, the sampling distributions for the Lorenz asymmetry coefficient and the Gini coefficient were based on 1000 replications. In addition, we provide a histogram of one of the 1,000 samples and a scatterplot of the two coefficients for each sample.

In all cases, the sampling distributions for both the Gini coefficient and the Lorenz asymmetry coefficient appear to be symmetric, even when the original distribution is highly skewed. In addition, there seems to be no relationship between these two coefficients, since the scatterplots appear to be circles. The one exception to this is the larger sample size for the lognormal distribution, in which there appears to be a very weak linear relationship between the two coefficients.

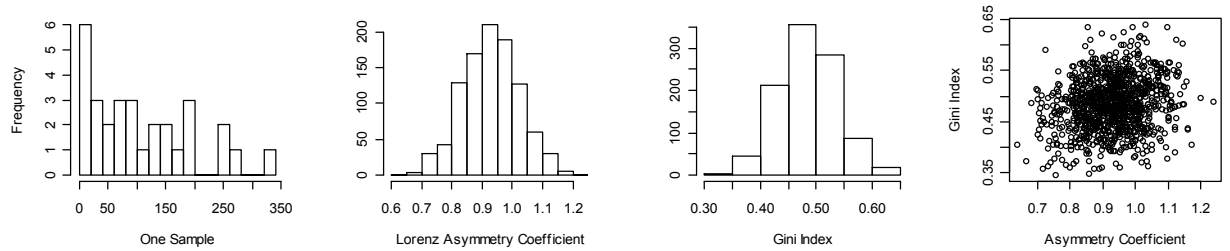
In another comparison of these two coefficients, we provide boxplots of their sampling distributions in Figure 3.20 and Figure 3.21. As expected, both coefficients have less variability when the sample size is larger. The Gini coefficient appears to better discriminate between the four highly distinct abundance distributions, since there is a greater separation in the Gini coefficient medians and less overlap in the sampling distributions. In contrast, the medians for the asymmetry coefficient are similar for all the distributions, and the overlap in the sampling distributions for the asymmetry coefficient would make it difficult for this coefficient to properly detect the underlying distribution.

The key results of this analysis are:

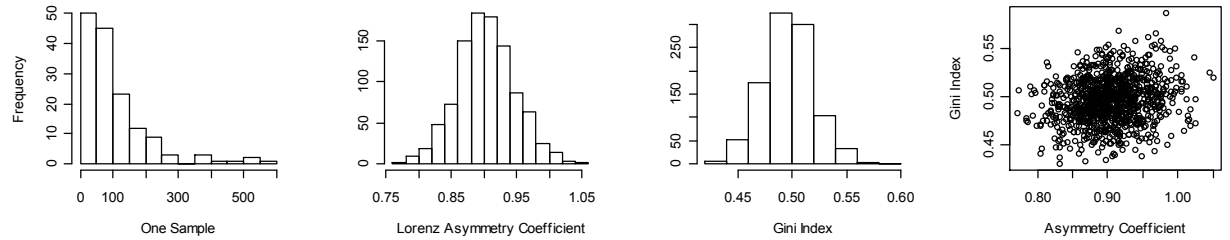
- The Gini index measures the inequity in a distribution. Its range is [0, 1], where 0 represents perfect equity and 1 represents perfect inequity
- The Lorenz asymmetry coefficient measures the symmetry of the Lorenz curve. Its range is [0, 2], where 1 represents perfect symmetry.
- Both the Gini index and the Lorenz asymmetry coefficient have symmetric sampling distributions.
- Simulation results indicate that the sampling distribution of the Gini index is more sensitive to the underlying distribution of counts, making it a better measure for distinguishing between unknown distributions.

Based on these preliminary results, we advocate continued exploration of the Gini coefficient as a test statistic for identifying rare and common OTUs, but we do not recommend further investigation into the Lorenz asymmetry coefficient.

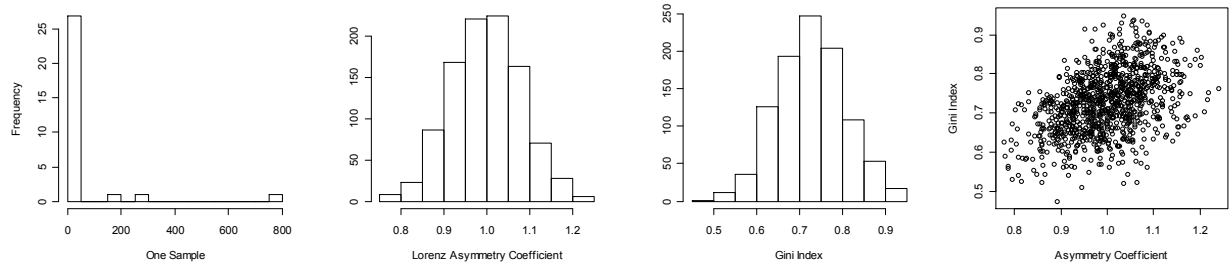
Exponential, sample size = 30



Exponential, sample size = 150



Lognormal, sample size = 30



Lognormal, sample size = 150

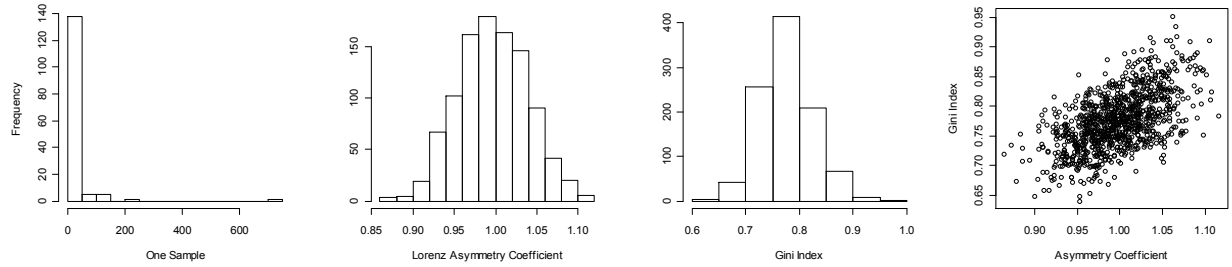
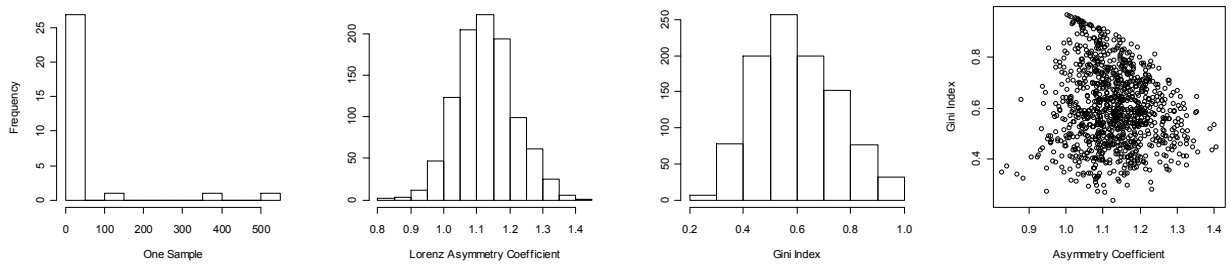


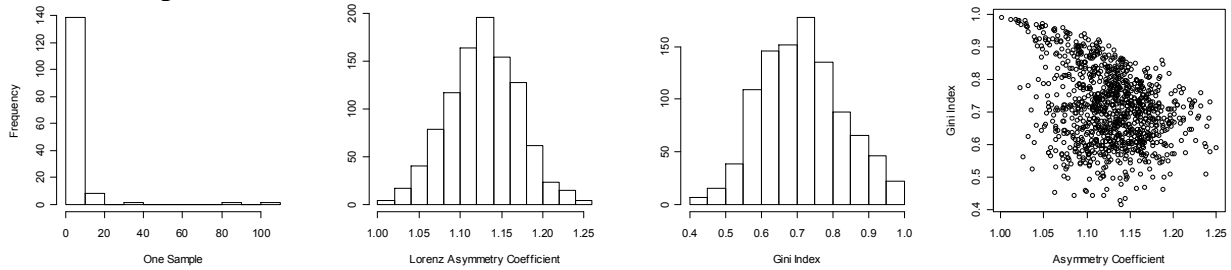
Figure 3.18: Sampling Distributions of the Gini and Asymmetry Coefficients. Part I

First panel shows the histogram of simulated abundances for <u>one</u> sample. The second and third panels show the sampling distribution for the Lorenz Asymmetry Coefficient and the Gini Coefficient, based on 1000 replicated samples. The last panel shows the relationship between the two coefficients.

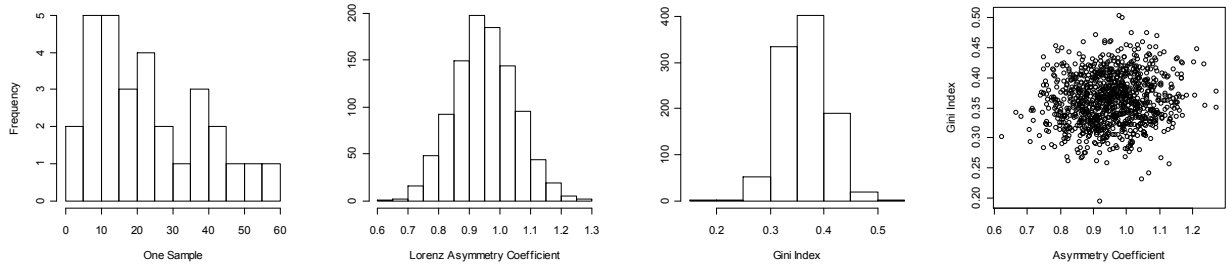
Pareto, sample size = 30



Pareto, sample size = 150



Gamma, sample size = 30



Gamma, sample size = 150

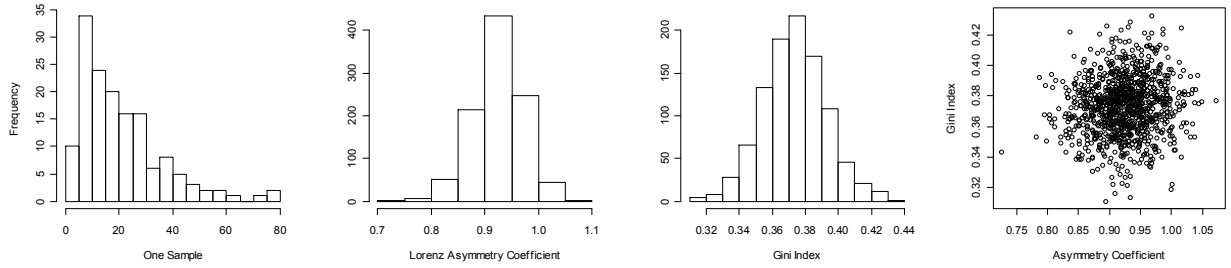


Figure 3.19: Sampling Distributions of the Gini and Asymmetry Coefficients. Part II

First panel shows the histogram of simulated abundances for <u>one</u> sample. The second and third panels show the sampling distribution for the Lorenz Asymmetry Coefficient and the Gini Coefficient, based on 1000 replicated samples. The last panel shows the relationship between the two coefficients.

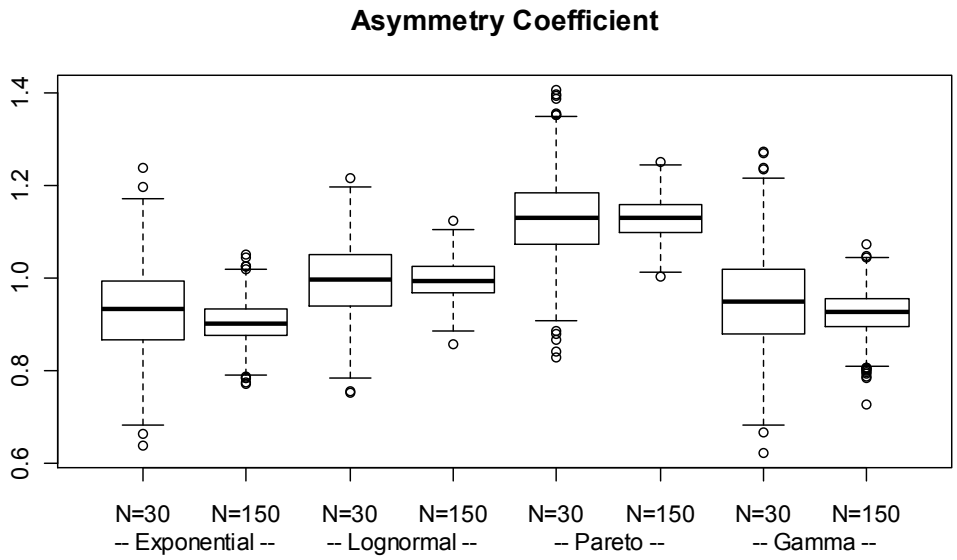


Figure 3.20: Comparative Boxplots for the Asymmetry Coefficient

The similarity of the medians and the overlap in the sampling distributions indicate that the Asymmetry Coefficient may be ineffective as a test statistic for distinguishing these distributions.

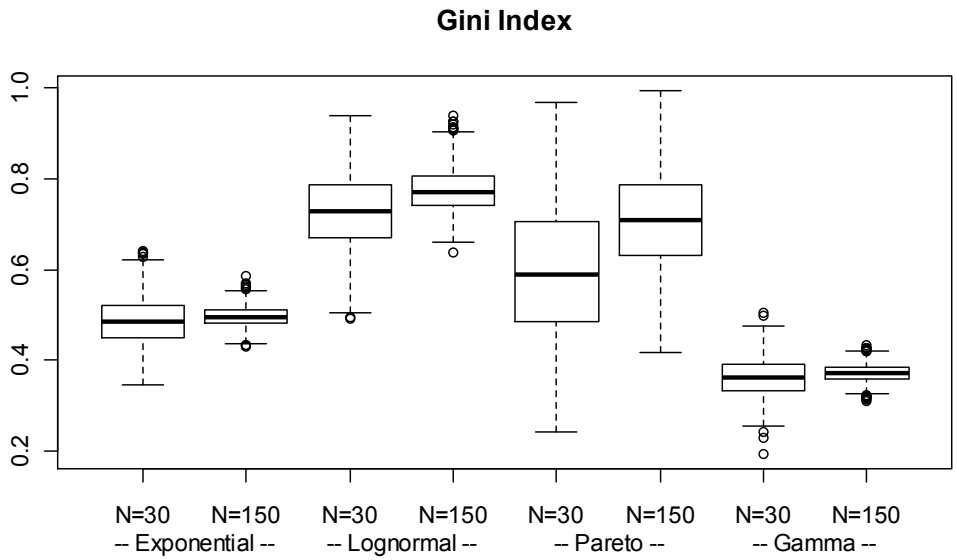


Figure 3.21: Comparative Boxplots for the Gini Coefficient

The sampling distributions for the Gini coefficient show greater separation is medians and less overlap among the various distributions. Thus the Gini coefficient could be a useful tool for discriminating between these distributions.

3.4. Probability Models

When we investigate possible distributions for OTU abundance data, we must consider the origins of the data. In particular, we are examining PCR-amplified segments of the DNA of microscopic organisms. The mechanisms that control the growth and decay of microscopic organisms in a natural environment are not yet well understood, since the technological advances necessary to monitor such organisms have only recently been attained. The counts that are recorded in an OTU data set are measuring the amount of microscopic organisms. In this setting, the concept of an 'individual' has no meaning, and thus the counts can be considered measurements from a continuous distribution, which are simply rounded to the closest integer.

For these reasons, we believe the most plausible distributions for OTU abundance data are continuous, nonnegative and strongly right-skewed. Such distributions include the exponential, gamma, lognormal and Pareto. From a mathematical perspective, the Pareto distribution is particularly attractive, since it provides a cohesive set of statistically plausible distributions for the individual entries in the OTU abundance matrix, and the corresponding column (OTU) totals. Specifically, let $X_i,\ i = 1, 2, \ldots, n,$ represent the nonzero abundances for *one* OTU at each of *n* sites. For this initial investigation, we assume the sites are independent, so that the X_i are independent. Under a null hypothesis that there is no difference among the sites, the X_i will also be identically distributed, so $\{X_1, X_2, \ldots, X_n\}$ form a random sample from some distribution *X*. Suppose $X \sim$ Pareto (1, *b*), that is, X has pdf $f(x) = b \cdot x^{-b-1},\ x \geq 1, b > 0$. Then $\log X \sim exponential(b)$ and $T = \sum_{i=1}^{n} \log X_i \sim gamma(n, b)$. When this is applied to an OTU data set, each OTU will have its own parameter *b*, which can be estimated from the data. Each OTU will also have its own value for *n*, the number of sites at which the OTU is present. Thus, for OTU *j*, $T_j = \sum_{i=1}^{n_j} \log X_{ij} \sim gamma(n_j, b_j)$. The *T*'s are *not* independent because the OTUs are not independent. Thus, under these assumptions, the precise form of the species abundance distribution is not known. But this is more coherent than the approach taken by some authors, for example, van der Gast *et al.* (2011), in which individual abundance values were presumed to follow a Poisson distribution, and their sum was fitted to a log series distribution.

There is another beneficial property of the Pareto distribution that may prove useful in detecting outliers in the individual abundance values, which in turn could illuminate outliers in

row and/or column totals. This property is defined in the following result, which is proven in Appendix C.

Result 3.4:

If $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ are the order statistics of a random sample from a Pareto(1,b) distribution, then the ratio of the largest to the second-largest $R = \frac{X_{(n)}}{X_{(n-1)}}$ also follows a Pareto (1,b) distribution.

Note that this implies that the mean and all the moments of the original distribution should match the mean and moments of the observed ratio $\frac{X_{(n)}}{X_{(n-1)}}$. This also implies that $\log R = \log X_{(n)} - \log X_{(n-1)}$ follows an exponential(b) distribution. This could be used to identify 'unusually large' values for individual abundances (*i.e.* values for one OTU at one site), which would be targeted for reduction in the algorithm described in Section 3.1.5.

Chapter 4. Methods and Results

4.1. Data Standardization

Pyrosequencing is a relatively new technology and, as with all new technologies, presents unique statistical challenges. Ongoing advancements in the technology continue to improve the quality of the resultant data, but there is still a large amount of variability in OTU data sets. Work reported in Chapter 3 described characteristics and challenges in OTU data and presented some initial approaches for summarizing and analyzing OTU data. The standardization technique described in this section builds from the common structure observed in the four OTU data sets. The proposed technique is a systematic and statistically defensible procedure to identify and reduce excess variability, that is, variability due to noise or technical artifacts resulting from the new technology. While it may seem mundane, it is necessary to preprocess any data set prior to performing statistical analysis. In fact, there is a large body of literature devoted to preprocessing of microarray data. The unique characteristics of pyrosequence data require unique preprocessing methods that have not yet been considered.

Individual entries in a pyrosequence data set represent the abundance, or count, for a particular OTU at a particular site. As noted in Section 3.1, approximately 90% of the entries are zero, and singletons can account for half or more of the nonzero values. We now consider the opposite extreme in the distribution of these counts, namely, the very large abundances. Each of the four data sets we have examined contain only a small number of extremely large individual counts, but they occur at least once in almost every site and are the major contributors to the total count at each site. In Lorena's data, for example, the two largest OTUs (with total counts 3949 and 3450, respectively) occur in every site, with individual counts ranging from 22 to 253. The third largest OTU, with total count 1261, occurs in 22 of the 37 sites, with individual counts ranging from 2 to 168. These three OTUs account for more than 50% of the total count in the data set, but less than 1% of all the OTUs and less than 5% of the individual nonzero values. They are among the most abundant OTUs in each site, and they dominate the smaller counts. This is illustrated in Figure 4.1

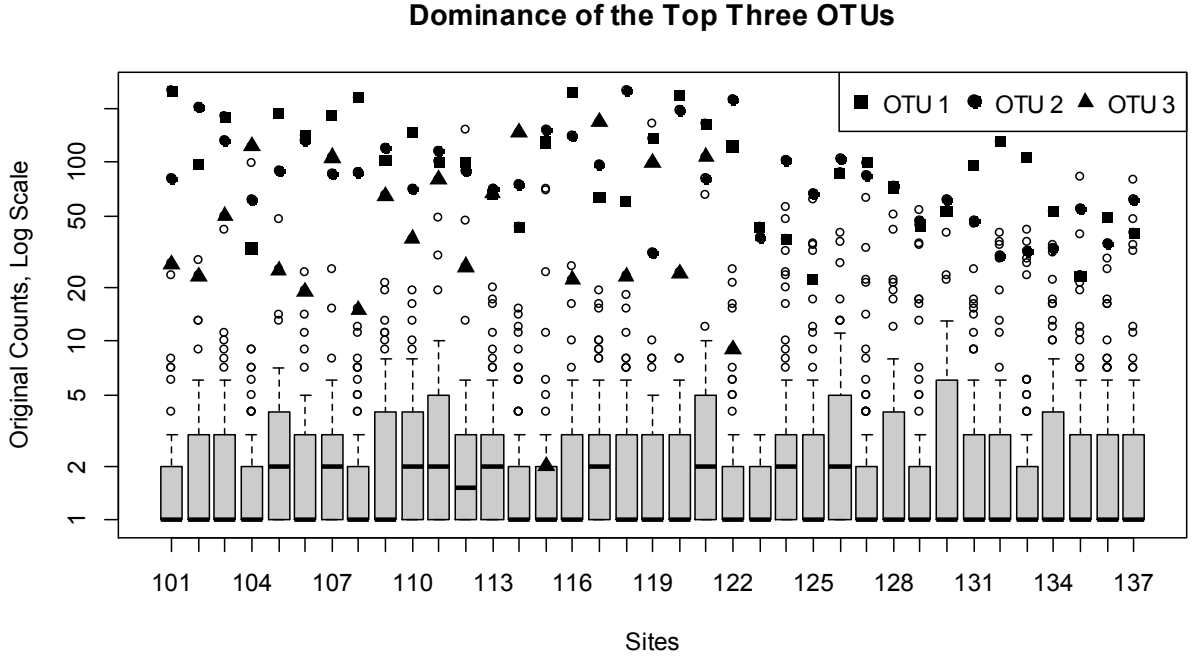


Figure 4.1: Dominance of the Three Largest OTUs in Lorena's Data

OTUs 1, 2 and 3 are among the most abundant OTUs in every site and comprise most the total count at each site. The presence of these large counts obscure the contributions of smaller counts.

Such large differences in scale among the OTUs will distort summary measures based on Euclidean distances, such as variance and correlation. One customary method for accommodating large differences is to center and scale each variable (OTU) across the sites. In ordinary circumstances, this is accomplished by subtracting some measure of location and dividing by some measure of variation. This is not feasible for pyrosequence data sets, in part because the extreme skew of the distribution inhibits reasonable and interpretable definitions of center and scale. Even when the zeros are disregarded, robust measures such as the median provide little useful information. In Lorena's data, for example, over three-fourths of all OTUs have median nonzero count equal to 1, and subtracting 1 will do little to center these highly skewed counts.

In Section 3.1.5, we presented an algorithm to reduce these overly large counts. This was a deterministic algorithm that forcibly reduced each 'large' count by a fixed percentage, while maintaining the rank order of the counts. This algorithm was applied to the soil data and reduced 7 individual counts in 7 different sites, but affected only two OTUs. Given the

prevalence of the three most abundant OTUs in Lorena's data, a more robust standardization method is needed.

Since the data are derived from a relatively new technology, sources of variability in the OTU datasets are currently not well understood. Variation can occur as a result of experimental and/or environmental conditions, and this is what we hope to discover as a result of the analysis. But variation can also occur as part of the data collection process, which was described in Section 1.2. For example, a small deviation in the concentration of amplicons in an analyte (a site) will affect all counts observed for that site, and primer bias can affect the amplification rate of specific OTUs, which can result in an entire column in the data matrix to be overly large or overly small. Such extraneous variability in the data needs to be identified before meaningful analysis can be conducted. Since amplification is a multiplicative process, we propose to use a multiplicative model to capture this excess variability. The results from this model can then be used to standardize the data.

4.1.1. Multiplicative Model

We propose to use a multiplicative model to standardize an OTU dataset. The model contains one parameter for each row (site) and one parameter for each column (OTU). Estimates of these parameters will capture the excess variability, which can be used to standardize the observed counts. This technique is borrowed from compositional data analysis, in which the data are centered prior to analysis (Daunis-I-Estadella, *et al*., 2006). To center a compositional data set, each value in the data matrix is divided by the geometric mean of the column. We propose to doubly-center the data set: first divide each entry by the geometric mean of the row, then divide the result by the geometric mean of the column. This procedure is very similar to one iteration of Tukey's median polish, with multiplication and division replacing addition and subtraction, and geometric means replacing medians. Median polish generates an additive model:

$$obs_{ij} = \mu + Row_i + Column_j + \varepsilon_{ij}$$

We are generating a multiplicative model:

$$obs_{ij} = (overall\ effect) \times \left(Row_i\ Effect\right) \times \left(Column_j\ Effect\right) \times e_{ij}$$

The overall effect is a scalar. For a data matrix containing N rows (sites) and D columns (OTUs), the row effect is a vector of length N, the column effect is a vector of length D, and the residuals are a matrix with dimension N×D.

The effects for the multiplicative model are estimated by iterating a multiplicative version of Tukey's median polish. Implementation of this procedure requires that all counts be strictly positive, thus the data matrix and the corresponding residuals matrix will contain missing entries in place of the zeros. In median polish, each iteration consists of a row sweep and a column sweep, where each sweep subtracts the row or column median. The algorithm has converged when each row and column median is nearly 0. The multiplicative version of this algorithm iteratively sweeps rows and columns, but each sweep consists of dividing by the geometric mean of the nonzero values in the row or column. The multiplicative algorithm converges when the geometric mean of every row and column is nearly 1. We call this procedure geopolish.

This algorithm was applied to Lorena's complete data, consisting of all 37 sites and all 799 OTUs. Unlike our earlier analyses, no sites and no OTUs were removed from the data. One of the sites (Plot 123) contains an extremely small number of OTUs and a low total count as compared to the other sites (see Figure 3.1). For all of our previous analyses, this site was deemed improperly amplified and was simply removed from the data. We keep this site in the current analysis, and examine the ability of the multiplicative model to detect that it is different. The data set also includes every OTU that was observed, even if it occurs in only one site and occurred as a singleton at that site.

Lorena's complete data set contains 2,578 nonzero counts and 1,450 singletons. The total count is 17,428 and the maximum individual count is 253. The large counts were *not* reduced prior to fitting the multiplicative model. Of the 799 OTUs, 407 occur in only one site and occur as a singleton at that site, 126 OTUs occur only as singletons, but in multiple sites, and an additional 47 occur in only one site, but with a count greater than 1. In most of our previous work these extremely rare OTUs were set aside, but we keep them in the current analysis.

The geopolish algorithm

1. Initialize the overall effect, all row effects and all column effects to be 1.
2. Calculate the geometric mean of the nonzero values for each row (site).
3. Divide each data value (individual count) by the geometric mean of its site.
4. Calculate the geometric mean of each column (OTU), using the nonzero (adjusted) values for the column.
5. Divide each adjusted value by the geometric mean of its OTU. These are the residuals.
6. Calculate the geometric mean of the values in Step 2.
7. Calculate the geometric mean of the values in Step 4.
8. Multiply the values in Steps 6 and 7. This is the overall effect for the current iteration.
9. Divide each value in Step 2 by the geometric mean from Step 6. These are the row effects for the current iteration
10. Divide each value in Step 4 by the geometric mean from Step 7. These are the column effects for the current iteration.
11. Multiply the overall effect from the previous iteration and the overall effect from the current iteration. This is the overall effect that will be carried into the next iteration.
12. For each row, multiply the row effect from the previous iteration and the row effect from the current iteration. These are the row effects that will be carried into the next iteration.
13. For each column, multiply the column effect from the previous iteration and the column effect from the current iteration. These are the column effects that will be carried into the next iteration.
14. Check the convergence criteria. If all of the geometric means calculated in Steps 2 and 4 are nearly 1, then stop. Otherwise, use the residuals from Step 5, the overall effect from Step 11, the row effects from Step 12 and the column effects from Step 13, and repeat Steps 2 through 14 until convergence.

Applied to Lorena's data, the geopolish algorithm converged after 38 iterations. A boxplot of the 37 estimated row effects shown in Figure 4.2 clearly indicates Plot 123 as an outlier, while the remaining row effects are centered at 1. The scatterplot indicates that the row effects are not highly correlated with the total count for the site.

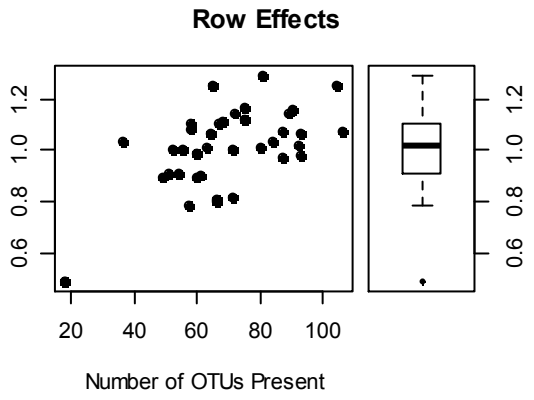


Figure 4.2: Estimated Row Effects versus number of OTUs present.
The one small site in Lorena's is clearly shown as an outlier, with estimated row effect 0.5. Row effects for the remaining sites are all near 1, indicating the overall similarity of these sites.

The boxplot of column effects, shown on a log scale in Figure 4.3, shows the extreme skew in the distribution. The column effects range from 0.6 to 71.9, where small effect sizes correspond to small, under-represented OTUs and large effect sizes correspond to more prominent OTUs. Since the data set was not trimmed to exclude the extremely rare OTUs, there are a large number of small column effects. A total of 540 OTUs (out of 799) have effect size less than 1, which indicates a 'below average' OTU. Unlike the row effects, the column effects do seem to be correlated with the total count for the OTU, when both are represented on a log scale. This is most likely due to the fact that the OTU total counts are highly skewed, with range [1, 3949] and median 1, and this extreme variation is captured by the column effects.

The geopolish algorithm generates one estimated row effect for each site and one estimated column effect for each OTU. These effects are, in essence, summary statistics of the corresponding sites and OTUs. These statistics can be used to test for differences between sites, or to classify OTUs as either rare or common. An example is given in Section 4.1.6. In order to perform these tests, we must have knowledge of the sampling distributions of these statistics. This is discussed in Section 4.1.4. In addition to the row and column effects, the geopolish algorithm generates one residual for each observed count in the dataset. These residuals are the standardized data. The distribution of the residuals is examined in Section 4.1.2 and examples of their interpretation are given in Section 4.1.5.

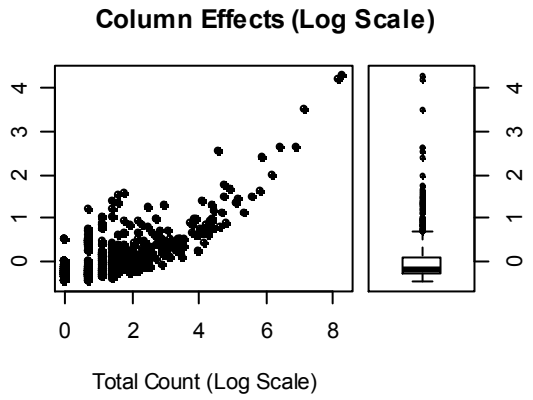


Figure 4.3: Estimated Column Effects versus total Count

There are many small column effects, corresponding to the many infrequent OTUs, but only a few large column effects for the more prominent OTUs. The vertical axis is the log of the column effects.

4.1.2. Residuals from the Multiplicative Model

We now explore the residuals from the multiplicative model. These are considered the standardized data, since they are free of excess variability in sites and OTUs as a result of the data collection process. Of particular interest is the distribution of the standardized data. If we can reasonably conclude that the distribution of the log-transformed standardized data is approximately normal, then usual parametric theory can be applied. On the other hand, if these data remain highly skewed, then nonparametric methods should be used to draw inferences.

A histogram of the standardized values are shown in the right panel of Figure 4.4 and the original counts are shown in the left panel. (Both histograms are given in log scale). The standardization has definitely changed the shape of the distribution. The skewness has diminished considerably and it is now centered at 0 (on a log scale). The excess of values in the interval (–0.5, 1] is of some concern, so we examine these residuals in more detail. Recall that the estimates for the multiplicative model were obtained using every nonzero count in Lorena's data, including all the singletons and the OTUs that occur at only one site. It is possible that these extremely rare OTUs and small counts are causing the spike in the histogram. We now explore some options to trim the data, and perhaps obtain a more symmetric distribution. We first consider removing the residuals associated with singleton counts.

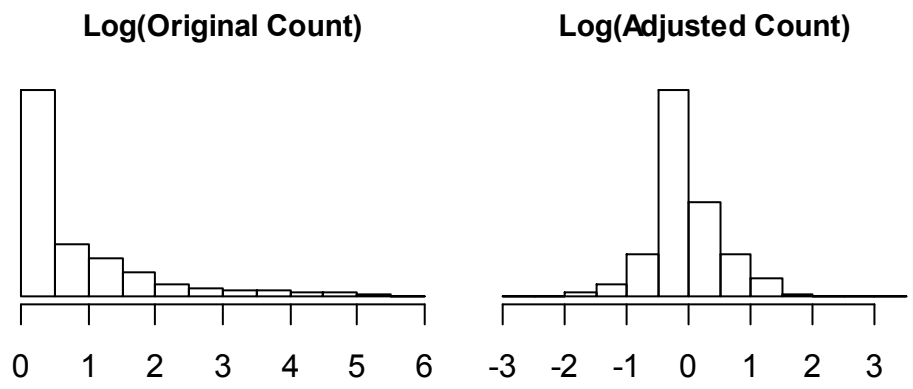


Figure 4.4: Distribution of Original and Adjusted Counts

When viewed on a log scale, the histogram of adjusted counts (i.e. residuals) from the multiplicative model (shown on the right) are centered at 0, while the original counts are clearly skewed.

The distribution of residuals is derived from all of the 2,578 nonzero counts present in the data, which includes 1,450 singletons. The residuals for the singletons are concentrated in the range (-0.5, 0] on a log scale. As shown in Figure 4.5, when these residuals are removed, the distribution of log residuals for the nonsingleton counts appears to follow a normal distribution, but the center is not at 0.

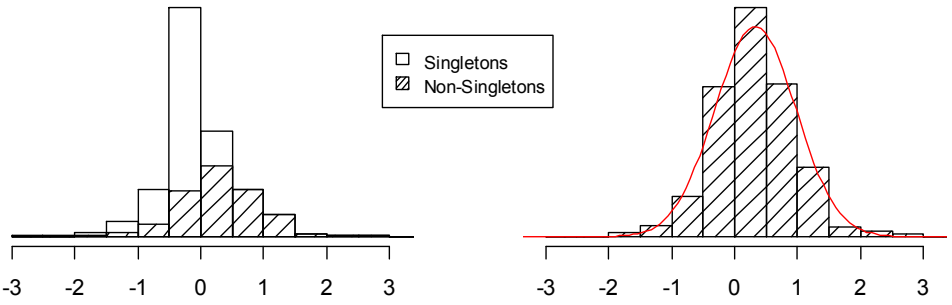


Figure 4.5: Distribution of Log Residuals

When viewed on a log scale, the histogram of residuals from the multiplicative model (shown on the left) are centered at 0, but are not symmetric with a spike just below zero. When the residuals associated with singleton counts are removed, the histogram appears symmetric (shown on the right), but the center is no longer at 0.

By visual inspection, it would seem that the logarithm of the residuals for nonsingleton counts follows a normal distribution, although each of five different tests for normality soundly

reject the normal distribution as a plausible model. These tests are Anderson-Darling, Cramer-von Mises, Lilliefors, Shapiro-Wilk and Pearson's chi-square, and their p-values range from 10^{-12} to10^{-6}. The normal QQ plot, shown in Figure 4.6, suggests that the deviation occurs in the tails, while a more detailed histogram indicates too much probability mass in the center of the distribution.

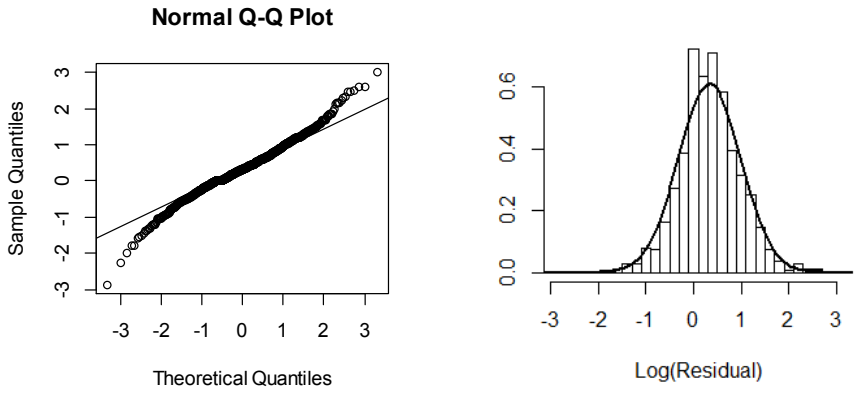


Figure 4.6: Normal Probability Plot and Histogram of Log Residuals

When viewed on a log scale, the histogram of residuals from the multiplicative model (shown on the left) are centered at 0, but are clearly asymmetric with a spike just below zero. When the residuals associated with singleton counts are removed, the histogram appears symmetric (shown on the right), but the center is no longer at 0.

It should be noted that all five normality tests are performed on all the nonsingleton residuals in the data set, so the sample size is 1,128. Such a large sample may be giving these tests too much power for detecting differences that are not of practical importance. Before we make a final decision regarding the normality of the log residuals, we consider trimming the data to reduce the excess number of residuals near the center of the distribution. We believe that the asymmetry in the distribution is caused, in part, by the singletons and extremely rare OTUs in the dataset. We consider the following seven options for trimming the data.

Option 1: Use all counts (this is the original data set with no trimming)

Option 2: Remove OTUs that occur in only one site, and occur as a singleton at that site.

Option 3: Remove all singletons from the data set. This may result in removing some OTUs from the data.

Option 4: Re-scale the vector of counts for each site by dividing by the total count for the site. This creates a compositional vector for each site, whose entries sum to 1.

Option 5: Remove all OTUs that either occur in only one site or occur only as singletons (perhaps in multiple sites). This will remove every column from the data set that has only one nonzero value (regardless of what the value is), and will also remove columns that contain only 1's.

Option 6: Remove OTUs that occur only as singletons, regardless of the number of sites in which the OTU is present.

Option 7: Remove OTUs that occur in only one site, regardless of the count for the OTU at that site.

The geopolish algorithm was performed on each trimmed data set. Summaries of the trimmed data sets are presented in Table 4.1, and histograms of the log residuals are given in Figure 4.7.

None of the data trimming options were successful at eliminating the asymmetry in the distribution of log residuals of all counts (histograms on the left in each panel of Figure 4.7), although most of the trimming options generated a symmetric distribution for nonsingleton counts (histograms on the right). It is interesting to note that trimming option 3, which removes all singletons, generated the most asymmetric distribution of log residuals. This seems to suggest that singleton counts are not the only source of skewed residuals.

Consider how the geopolish algorithm generates the residuals. In particular, we focus on the steps involving the column sweeps and how these sweeps affect a column that contains only one positive count. These columns represent OTUs that occur in only one site. The geometric mean for this column is equal to the lone nonzero count, so when this count is divided by the geometric mean the result (the residual) is always equal to 1, within round-off error. Thus all of the effect for this OTU is swept into column effect, leaving no variability in the residual. This occurs for 454 of the 799 OTUs in Lorena's original data set, and these counts comprise nearly 18% of the nonzero individual counts in the data. All of these counts are removed under options 5 and 7, and some are removed under options 2, 3 and 6. The effect on the distribution of the log residuals is shown in Figure 4.8. All of these histograms exhibit asymmetry with an excessive number of log residuals slightly below 0.

	Option 1	Option 2	Option 3	Option 4	Option 5	Option 6	Option 7
Number of Sites	37	37	37	37	37	37	37
Number of OTUs	799	392	266	799	219	266	345
Number of Nonzero Counts	2578	2171	1128	2578	1790	1837	2124
Number of Singletons	1450	1043	0	0	709	709	1043
Total Count	17,428	17,021	15,978	37	16,572	16,687	16,906
Range of Individual Counts	[1, 253]	[1, 253]	[2, 253]	[0, 0.52]	[1, 253]	[1, 253]	[1, 253]
Range of Site Totals	[104, 589]	[99, 578]	[93, 548]	[1, 1]	[95, 557]	[97, 566]	[97, 569]
Range of OTU Totals	[1, 3949]	[2, 3949]	[2, 3949]	[0, 8.38]	[3,3949]	[2, 3949]	[2,3949]
Range of Number of OTUs at a Site	[18, 106]	[13, 83]	[7, 53]	[18, 106]	[10, 68]	[11, 71]	[12, 80]
Range of Number of Sites for an OTU	[1, 37]	[1, 37]	[1, 37]	[1, 37]	[2, 37]	[1, 37]	[2, 37]
Converged at Iteration	38	30	32	38	26	27	30

Table 4.1: Impact of Data Trimming Options on Lorena's Data

The data trimming options are designed to reduce the number of small counts, and the impact is most dramatic in the reduction of number of OTUs. Option 4 transforms each site vector into a compositional vector, so the values represent proportions. The remaining options show little impact on either the total count or range of counts in the data set.

The residual analysis indicates that any OTU with only one positive count will have a log residual equal to 0, within roundoff. These counts, among others, contribute to the asymmetric spike in the center of the distribution. Even when these counts are removed, the distribution of log residuals is still asymmetric with an excessive number of values slightly below zero. Some of these residuals are associated with singleton counts, but they are also associated with nonsingleton counts and with OTUs with a wide range of site occurrences and total counts. At the present time, it seems the asymmetry in the residuals is a result of the asymmetry in the original counts and the unknown dependencies between OTUs.

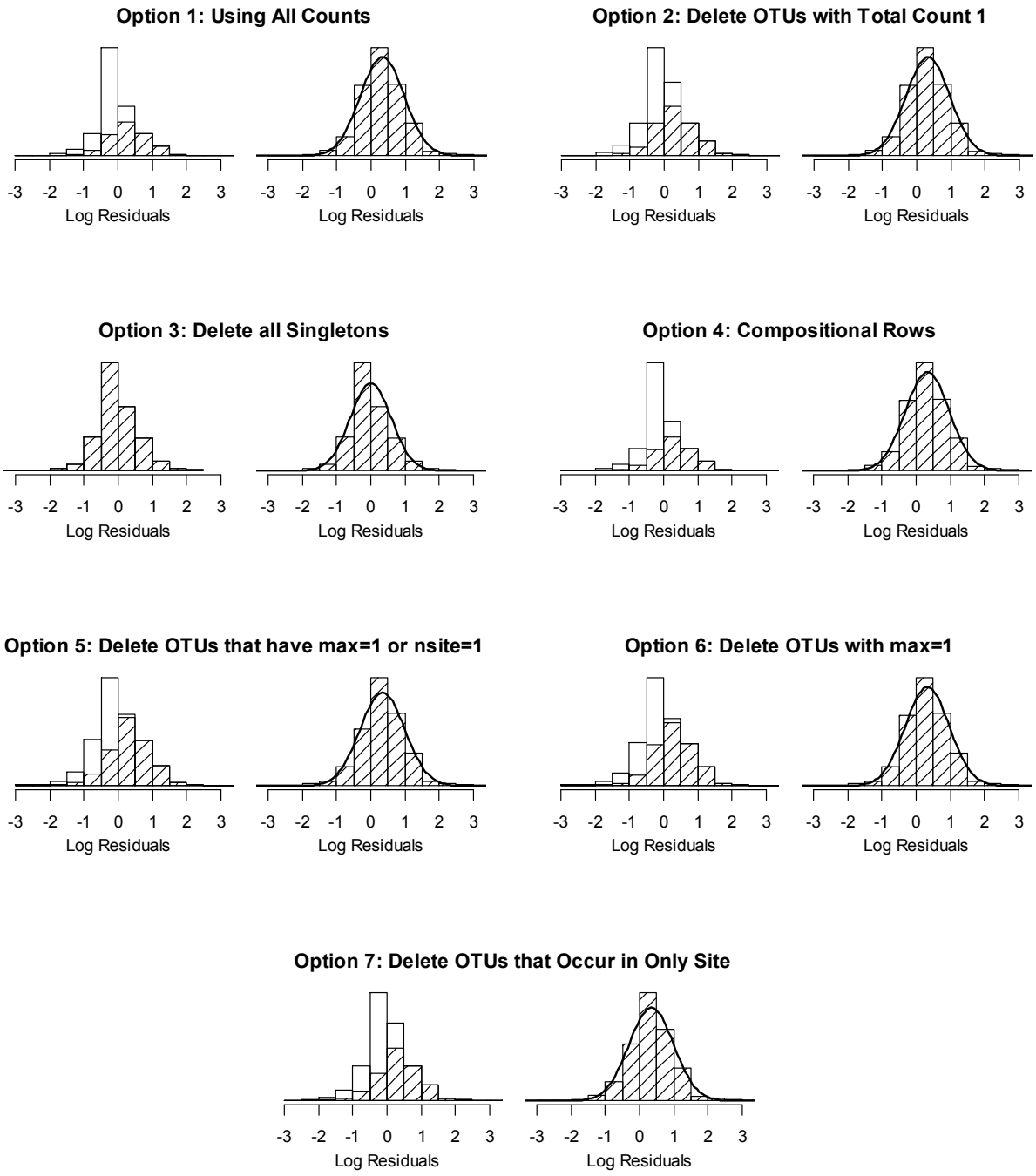


Figure 4.7: Distribution of Log Residuals Under Various Data Trimming Options

Lorena's data was trimmed according to the seven options and each resulting dataset was fit to the multiplicative model. In each panel, the histogram on the left shows the distribution of all log residuals and the histogram on the right shows the distribution of log residuals that are associated with nonsingleton counts.

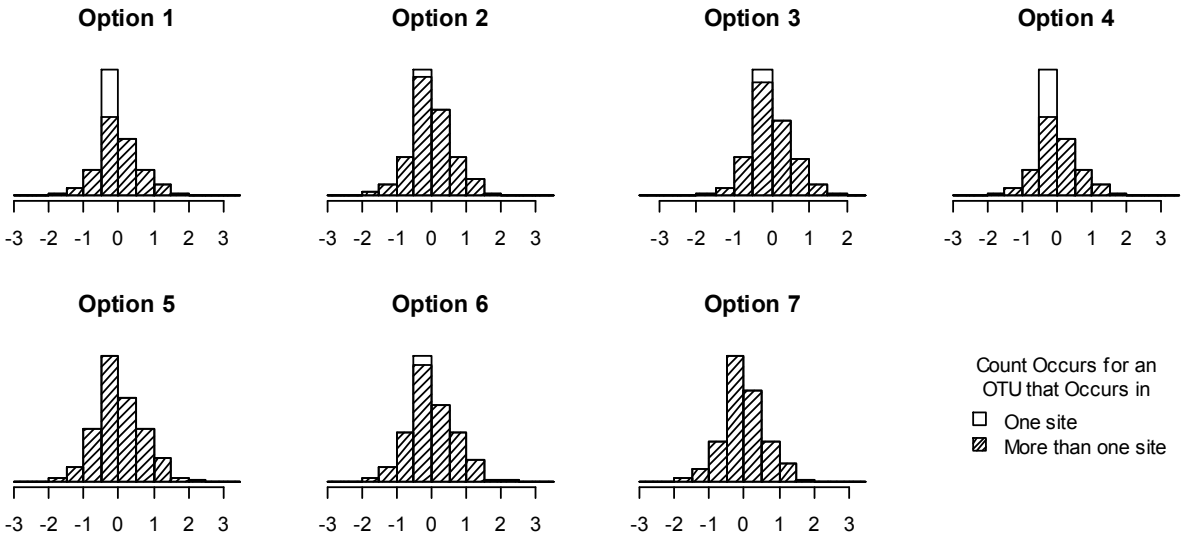


Figure 4.8: Distribution of Log Residuals, Showing Solo Counts for OTUs

Log residuals arising from OTUs that occur in only one site are shown in white. Some trimming options removed all of these counts and their associated residuals. The spike near the center of each histogram remains intact.

Since none of the data trimming options were successful at removing the spike near the center of the distribution of residuals, there is no motivation for trimming the data prior to fitting the multiplicative model. We will therefore continue to use Lorena's full data set, but we will be cautious in using inference that is sensitive to the assumption of normal, independent errors. The multiplicative model captures much of the variability in sites and OTUs that can be attributed to the data collection process. The residuals from this model are the adjusted, or standardized, counts. When viewed on a log scale, the adjusted counts have a mound-shaped distribution, instead of the extremely skewed distribution of the original log counts. Furthermore, the multiplicative model is robust to extremely small counts and infrequent OTUs, so all of the data can be standardized and used in subsequent analysis.

Although the usual assumption of independent, normal errors in linear models may not be satisfied, the structure of the multiplicative model is equivalent to a log linear model. In the next section, we explore this relationship in order to compare the estimates obtained from the geopolish algorithm to ordinary least squares estimates.

4.1.3. Relation to Ordinary Least Squares

The estimates generated by geopolish algorithm are equivalent to ordinary least squares estimates of a log-transformed model. The multiplicative model is

$$x_{ij} = a \times R_i \times C_j \times e_{ij}, i = 1, \ldots N, \; j = 1, \ldots D$$

where

x_{ij} is the observed count for OTU j at site i

a is the overall effect

R_i is the row effect for site i

C_j is the column effect for OTU j

e_{ij} is the random error

This is equivalent to the log linear model

$$y_{ij} = \alpha + \beta_i + \gamma_j + \varepsilon_{ij}$$

where $y_{ij} = \log(x_{ij})$ and the $\beta's, \; \gamma's, \; \varepsilon's$ are the row effects, column effects, and errors on a log scale.

We fit the log linear model to Lorena's data containing all counts (so nothing was trimmed), using the R function `lm`. The diagnostic plots, shown in Figure 4.9, indicate that the assumption of independent normal errors may not hold. The location-scale plot also reveals repetitive V-shaped patterns, which are produced by the many small duplicate counts in the data. We now compare the diagnostic plots from the log linear model to those obtained from the multiplicative model. To facilitate comparison, we log-transform the results from the multiplicative model. The two diagnostic plots from the multiplicative look very similar to those from the log linear model, and in fact, they are identical. When viewed on a log scale, the estimated multiplicative model is exactly equal to the estimated log linear model. This is not a coincidence. It can, in fact, be shown that this is true in general. The derivation is straightforward, but tedious, and is not included here.

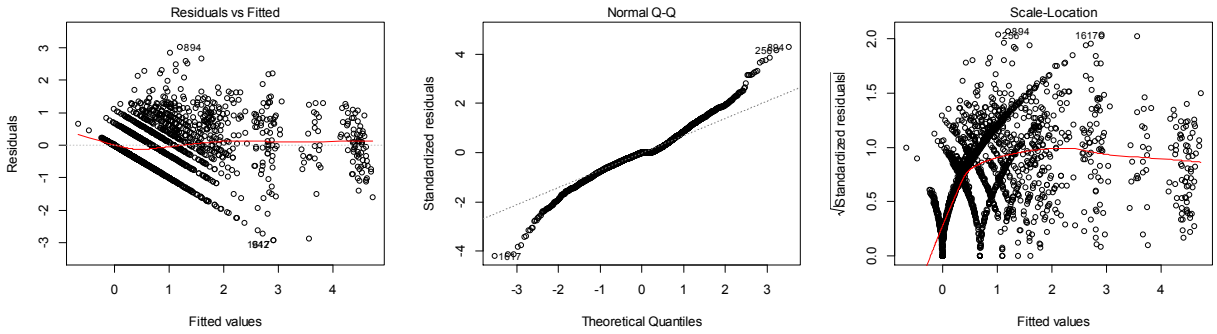


Figure 4.9: Diagnostic Plots for the Log Linear Model

The distinctive patterns in the two plots on the left and the curvature in the normal probability plot are indications that the errors may not be independent and normally distributed.

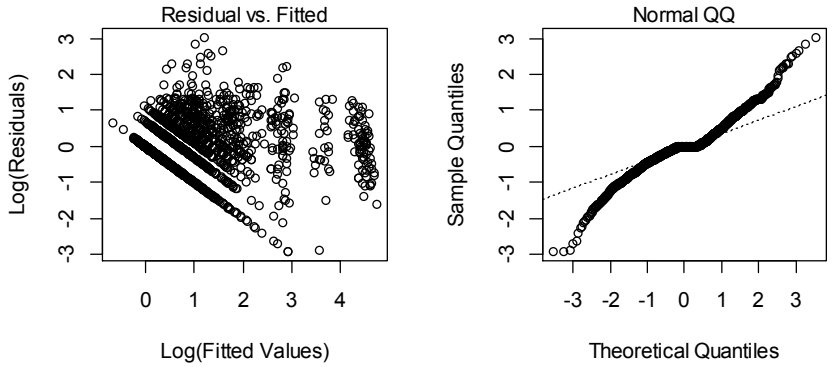


Figure 4.10: Diagnostic Plots (Log Scale) for the Multiplicative Model

The distinctive patterns in the two plots on the left and the curvature in the normal probability plot are indications that the errors may not be independent and normally distributed.

The multiplicative model generates an estimated effect for every row and every column, while the log linear model treats rows and columns as factors, and defines appropriate indicator variables for the levels of each factor. The resulting model matrix is not of full rank, and additional constraints must be placed on the parameters in order to identify a unique least squares solution. One common constraint is to select one level of each factor as the reference level and set its parameter equal to zero. Then the estimates for the remaining levels of this factor are in relation to the reference level. Another common constraint is to require that the estimated coefficients for each factor sum to zero. We will show that the log-transformed estimates from

the multiplicative model also satisfy this constraint, and that the parameter estimates from the two models are equal.

The solution found by the geopolish algorithm satisfies the log linear least squares estimating equations, and therefore is a least squares solution. This is a direct result of the convergence criteria and the normalization that occurs within the algorithm. The convergence criteria requires that, for each row and each column, the product of the residuals will equal one. Therefore, the sum of the log residuals for each row and each column will equal zero. The normalization that occurs in Steps 9 and 10 of the algorithm guarantee that the product of the row effects will equal one and that the product of the column effects will also equal one, thus the sum of the log of these effects will equal 0. Therefore, log-transformed estimates from the multiplicative model satisfy the least squares estimating equations and the solution is a least squares solution. In addition, the sum of the log coefficients for each factor (row and column) sum to 0, so the least squares solution found by the geopolish algorithm is the same as that found by OLS in the log linear model.

4.1.4. Standard Errors of the Estimates

Although the geopolish algorithm is able to estimate every model parameter, it does not provide estimates of their standard errors. Information about the standard errors is required for any form of inference regarding these estimates. Least squares theory provides standard errors of the estimates of the parameters in the log linear model, and we can apply the Delta Rule to adjust the known standard error for the log transformation. The relationships between the parameters are

$$\alpha = \log(a);\ \beta_i = \log(R_i);\ \text{and } \gamma_j = \log(C_j), \text{ or}$$

$$a = \exp(\alpha);\ R_i = \exp(\beta_i);\ \text{and } C_j = \exp(\gamma_j)$$

so, by the Delta rule,

$$se(\hat{a}) = se(\hat{\alpha})\exp(\hat{\alpha});\ se(R_i) = se(\hat{\beta}_i)\exp(\hat{\beta}_i);\ \text{and } se(\hat{C}_j) = se(\hat{\gamma}_j)\exp(\hat{\gamma}_j).$$

This approach has two major drawbacks. First, the Delta Rule is merely a large sample approximation and uses normal distribution theory. Thus it may not provide accurate results for some parameters. Second, it relies on the accuracy of the standard errors generated by the log linear model. As we will show below, the lack of compliance with the log linear model

assumptions generates standard errors that are too large. Therefore, these inflated standard errors should not be used to generate standard errors for the multiplicative model.

The key to understanding why the log linear standard errors are too large lies in the histogram of log residuals from the multiplicative model, which are the same as the residuals from the log linear model. This was shown in Figure 4.5 and is reproduced here (Figure 4.11) for convenience. The assumption of normal errors is not valid because there are too many residuals in the interval (–0.5, 0]. Thus the actual residuals are more compressed toward zero than modeled by the normal assumption. As a result, the log linear model will produce standard errors that are too large.

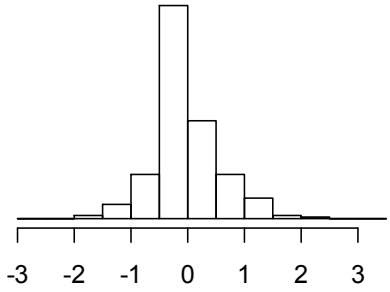


Figure 4.11: Log Residuals from the Multiplicative Model
These are also the residuals from the log linear model.

To obtain accurate standard errors for the multiplicative model estimates, we will bootstrap the residuals from that model. These residuals are free from both row and column effects, so that under a null hypothesis of 'no difference' (between treatment and/or environmental conditions), we can argue that each residual is equally likely to be observed with any row or column. By permuting the residuals, we are we are assigning them to a new row and column combination. We then multiply the original estimates for the row, column and overall effects with the newly permuted residual, and obtain a new dataset from which we can generate new estimates of the row, column and overall effects. By repeating this process numerous times, we can estimate the sampling variability of these estimated effects. This technique is analogous to the residual bootstrap for linear models (*cf.* Efron and Tibshirani, 1994).

By resampling the original residuals, we are preserving the concentration of residuals near 1 (that is, the log residuals near 0), as shown in Figure 4.11. This produces standard errors that are smaller than those generated by the log linear model, which erroneously assumes a more

dispersed distribution of residuals. For 1,000 resamples, the differences between the estimated standard errors for site effects are shown in the left panel of Figure 4.12, and standard errors for OTU effects are in the right panel. The range of estimated OTU effects is large (0.6 to 71.9), and this is reflected in the standard errors. In contrast, the range of estimated site effects is only 0.5 to 1.3, and the difference between the two sources of standard errors appears to be multiplicative. For the site effects, the slope of the regression line through these points is 0.833 (se = .003), with correlation 0.988.

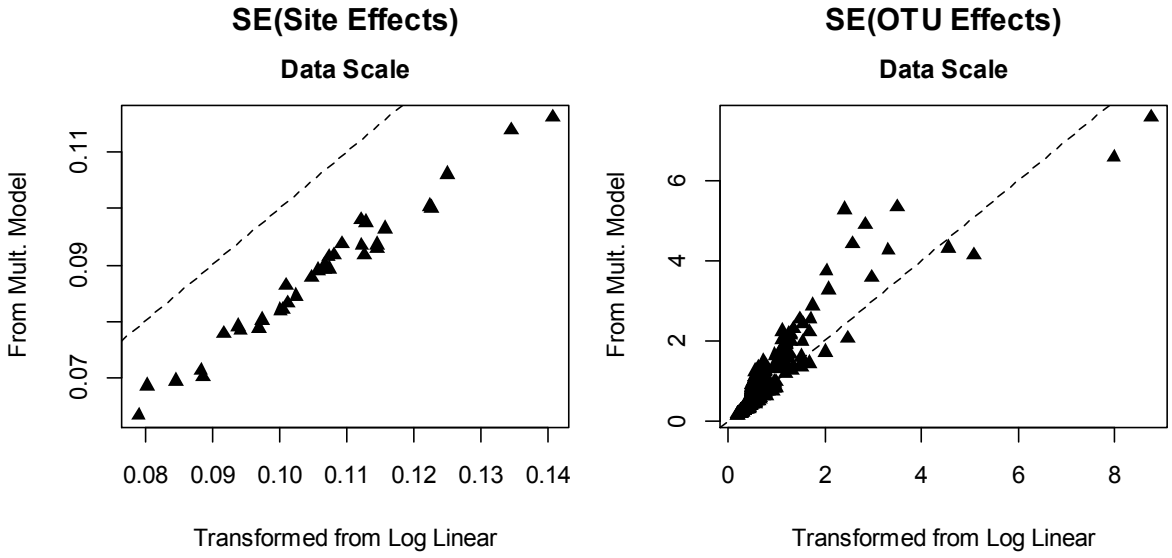


Figure 4.12: Standard Errors for Site and OTU Effects

The y-axis represents the estimates from resampling and the x-axis are the estimates obtained from the log linear model, transformed via the Delta rule. The dashed line represents equality between the two estimates.

4.1.5. Advantages of the Multiplicative Model

Other methods to model or analyze OTU data may require the removal of some rows and some columns from the dataset. Sites (rows) with low total count are often removed because the sites are deemed improperly amplified, and OTUs (columns) with many zeros generate numerically unstable estimates for some summary statistics. In contrast, the multiplicative model is robust for these extreme data and generates estimated effects for every site and every OTU in the data set. The row effects, column effects and residuals estimated from the multiplicative model can be used in various ways, as both summary statistics and to standardize

the observed counts. In particular, the residuals can be interpreted as the standardized counts, less influenced by the variability associated with the data collection process. Examination of these adjusted counts often reveal an underlying pattern in the data that is obscured in the original counts.

Consider the counts for Site 112, as shown in Figure 4.13. As expected, many OTUs have very small counts, but OTUs 3 and 31 have larger counts, and all of these are dominated by the counts for OTUs 1, 2 and 6. It is known that OTUs 1, 2 and 3 have large counts in nearly every site, but we do not know if the counts observed in Site 112 are typical *for these OTUs*.

When we consider the adjusted counts, a very different pattern of abundances emerges, as shown by the open triangles in Figure 4.14. The adjusted counts indicate where the observed counts fit in relation to all the counts for the OTU. A small adjusted count indicates that the observed is small *for this OTU*, an adjusted count near 1 indicates that the observed count is fairly typical for the OTU, while values greater than 1 indicate a large count for this OTU. The adjusted counts shown in Figure 4.14 reveal that the counts for OTUs 6 and 31 are unusually large, while the counts for OTU 1, 2 and 3 are fairly typical.

The distribution of original counts for these OTUs, shown in Figure 4.15, verify these conclusions. Even though the counts for OTUs 1, 2 and 3 are among the largest in Site 112, these OTUs have large counts in many sites. Their counts in Site 112 are neither unusually large nor unusually small, so their adjusted counts are near 1. The largest count in Site 112 occurs for OTU 6, and this is the second-largest count (in any site) for this OTU. This is an unusually large count, which is reflected in the large adjusted count (greater than 10). The count for OTU 31 is even more unusual, even though it is only 47. This is the largest count for this OTU, and its next-largest count (in any site) is only 7. Its adjusted count is slightly more than 12.

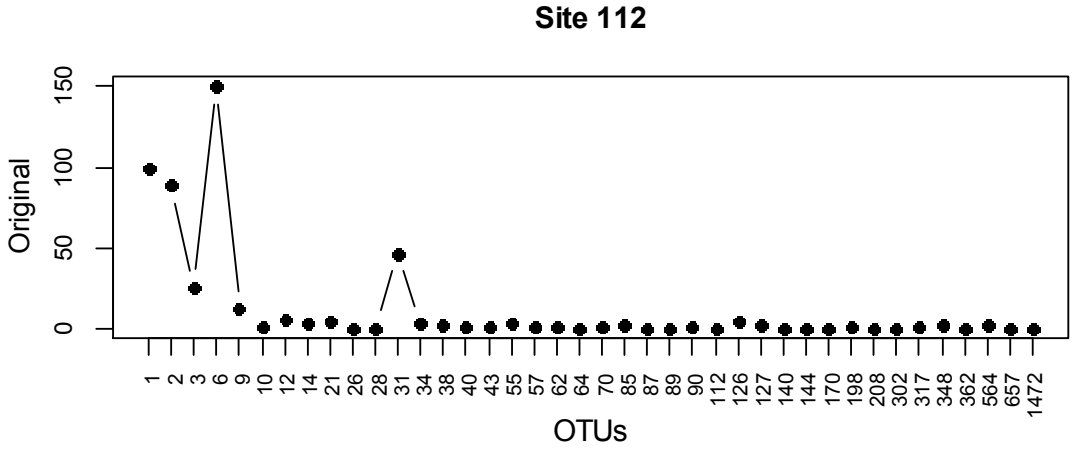


Figure 4.13: Observed Nonzero Counts at Site 112

Site 112 contains 52 unique OTUs, but 26 occur as singletons. Some of the singleton counts are not shown. OTU 6 has the largest count (150), followed by OTUs 1, 2, 31 and 3, with counts 99, 89, 47 and 26, respectively.

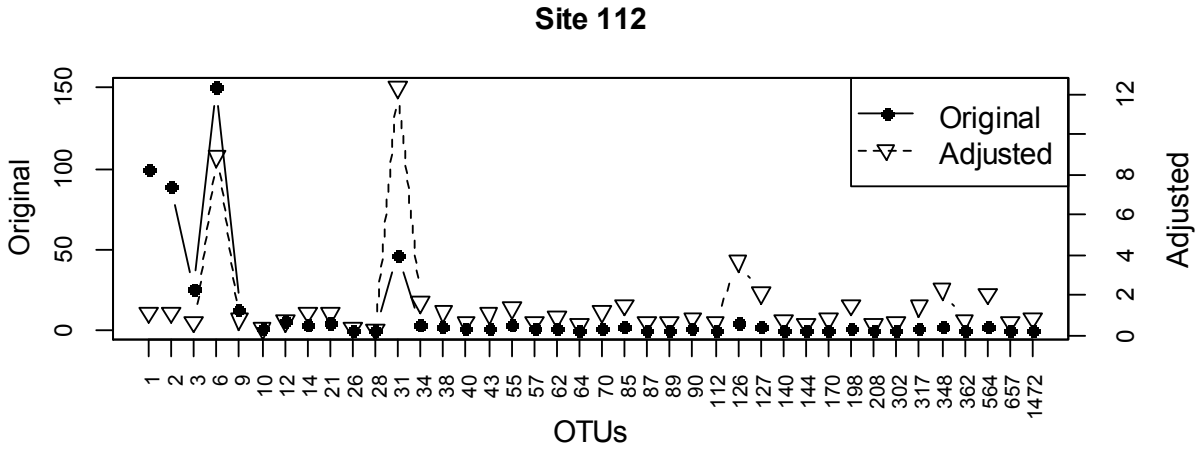


Figure 4.14: Adjusted Counts at Site 112

The adjusted counts reveal that the counts for OTUs 31 and 6 are unusually large for these OTUs. In contrast, while the original counts for OTUs 1, 2 and 3 are large, these OTUs have many large counts (in other sites), so the counts observed in Site 112 are not considered large for these OTUs. Note the change in scale on the two y-axes. The largest original count is 150 (for OTU 6), while the largest adjusted count is only 12 (for OTU 31).

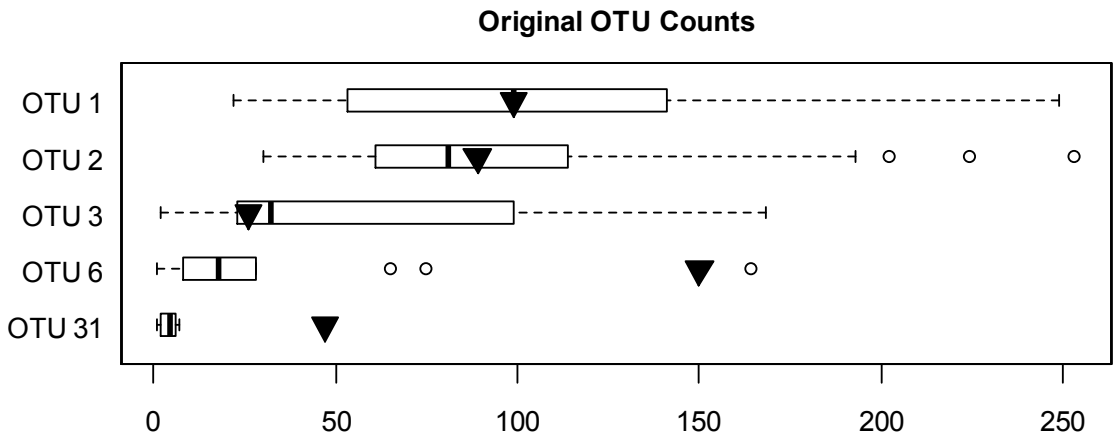


Figure 4.15: Distribution of All Observed Counts for Five OTUs

These boxplots show the distribution of all observed counts (across all sites) for the five most abundant OTUs in Site 112. The large triangle is the count in Site 112. The counts for OTUs 1, 2 and 3 are fairly typical for these OTUs, while the counts for OTUs 6 and 31 are unusually large.

4.1.6. Testing for Differences Across Sites

Lorena's dataset contains a descriptive variable for each site, corresponding to the area in Bolivia (Ancoraimes or Umala) in which the site is located. Of the 37 sites, 23 are in the Umala region and 14 are in Ancoraimes. We want to know if there is a difference between Ancoraimes and Umala sites. Depending on the research objective, it may be desirable to test specific OTUs for differences between the sites. If the hypothesis involves only one OTU then the original counts can be compared directly, that is, adjusted counts are not necessary. On the other hand, if we want to compare sites using a collection of OTUs at each site, then we should consider using the adjusted counts. The adjusted counts can be used to calculate summary statistics for each site, and these summary statistics can serve as the response variables on which we base the test. An alternative approach is to use the row effects themselves as the response variable. We illustrate these approaches using Lorena's data, testing for differences between Umala and Ancoraimes sites.

We consider several univariate measures to summarize each site, and compare sites on the basis of these measures. One of the most common univariate summary statistics used in ecological studies is the diversity measure Fisher's alpha (see Appendix B), but there are other choices. Possibilities include the Gini index calculated for each site (as a measure of

unevenness), the total count for a site, and the row effects from the multiplicative model. For both the total count and the Gini index, we also have a choice to use the original counts or the adjusted counts. However, estimation of Fisher's alpha requires all values to be integers, so the original counts must be used. Boxplots of these measures are shown in Figure 4.16.

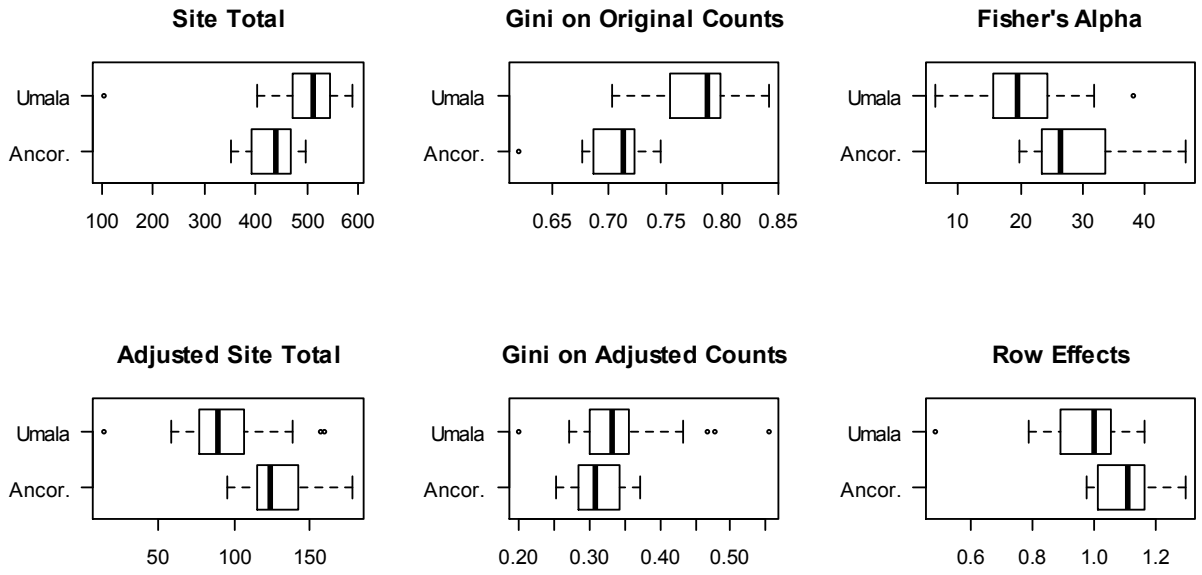


Figure 4.16: Boxplots of Summary Measures for the Sites

These measures can be used as response variables when testing for differences between Umala and Ancoraimes sites.

The boxplots of the site totals and the Gini indexes are very different for the original counts than they are for the adjusted counts. This is because the column effects vary from 0.6 to 71.9, with larger values for more abundant OTUs. When the counts are adjusted, the large counts for abundant OTUs are, in essence, divided by a large column effect. While the original counts range from 1 to 253, the adjusted counts are all between 0.05 to 20.69. By reducing the magnitude and spread of these values, the Gini index interprets the adjusted counts as more evenly distributed, and therefore the Gini values are closer to 0. These changes affect the Umala sites more than the Ancoraimes sites because the Umala sites contain most of the large counts in the data set. Of the 40 largest counts, 35 are from Umala sites.

As an aside, the boxplots in Figure 4.16 indicate a striking similarity in the distributions of Fisher's alpha and the adjusted site totals. These two measures are highly correlated (0.87), as

indicated in Figure 4.17. At present, it is unknown if there is a quantifiable relationship between these two measures, or if this is merely a coincidence in Lorena's data.

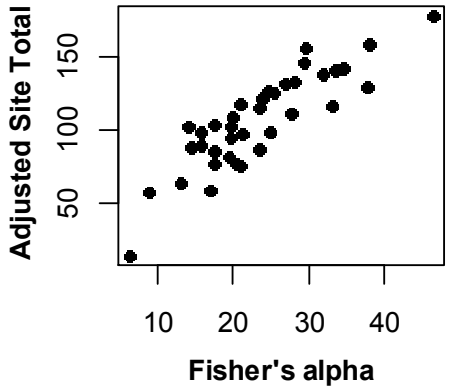


Figure 4.17: Relationship between Fisher's alpha and the Adjusted Site Totals
These measures can be used as response variables when testing for differences between Umala and Ancoraimes sites.

Each of these six summaries is used as a response variable to test for differences between Umala and Ancoraimes sites. Since the boxplots indicate asymmetric distributions with outliers, we employ the Kruskal-Wallis rank-based test. The p-values, shown in Table 4.2, indicate that all tests are significant at $\alpha = 0.05$, with the exception of the Gini index calculated from adjusted counts. For comparison, we also performed these tests using ordinary ANOVA. Although the QQ plots (Figure 4.18) show troublesome departures from normality, the p-values are all similar to those from the Kruskal-Wallis test.

Response variable	p-values	
	Kruskal-Wallis	ANOVA F
Original site totals	0.0007	0.0346
Adjusted site totals	0.0418	0.0584
Original Gini index	4.8e-06	0
Adjusted Gini index	0.1497	0.1229
Fisher's alpha	0.0013	0.0016
Row effects	0.0011	0.0008

Table 4.2: P-values for Testing Umala vs. Ancoraimes Sites
With the exception of the Gini index calculated from adjusted counts, all Kruskal-Wallis tests are significant at $\alpha = 0.05$. Results from ANOVA are provided solely for comparison; we should rely on the results from the Kruskal-Wallis tests.

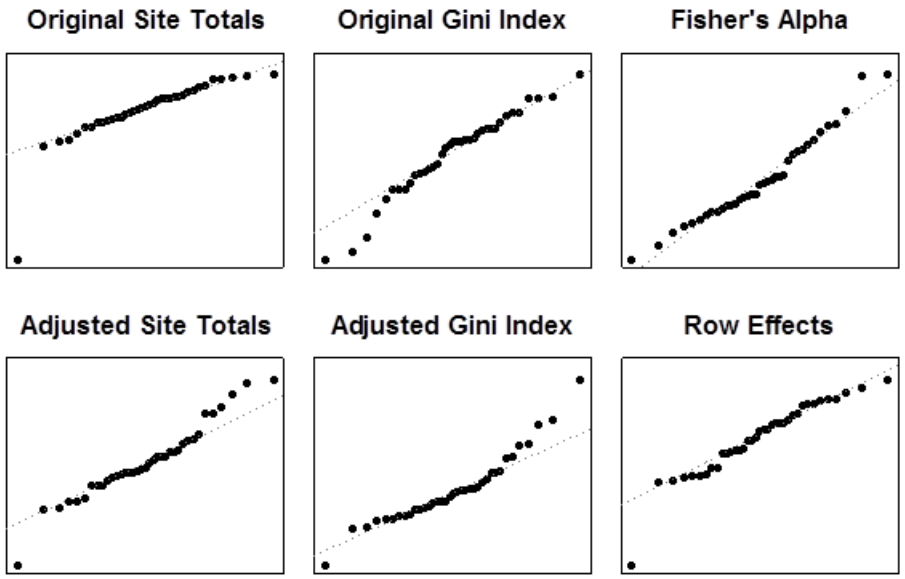


Figure 4.18: Normal Probability Plots from ANOVA

Outliers and distinct curvature indicate the assumption of independent normal errors may be not satisfied. We should base our decision on the nonparametric Kruskal-Wallis test.

We now explore a second method for utilizing the row effects from the multiplicative model in order to detect differences between Umala and Ancoraimes sites. We divide the dataset *by rows* into two groups, according to the two geographic regions. Separate estimates are obtained for the multiplicative model: once using only Umala sites and once using only Ancoraimes sites. These estimates are then compared to the estimates obtained using the full data set.

The results, shown in Figure 4.19, clearly show a separation between regions when only a subset of the data is used to estimate the site effects. Ancoraimes sites have larger site effects when the full data set is used, while site effects for Umala sites are smaller when the full data set is used. This pattern of separation may be useful in detecting differences in OTU compositions between the two types of sites, but at this time it is not clear how to construct a viable test statistic that incorporates this information.

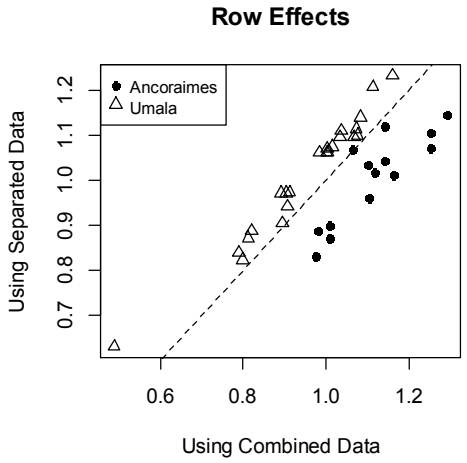


Figure 4.19: Comparing Row Effects Using Subsets of Data

There is a very distinct separation between the site effects for Umala and Ancoraimes regions when only the data from the region are used to estimate the effects.

4.2. Theoretical Results for the Gini Index

The Gini index was introduced in Section 3.3 as an alternate measure for classifying rare and common OTUs. We advocate its use as a replacement of the 50% persistence threshold criterion proposed by Magurran and Henderson (2003). In this section, we present some properties of the Gini index that make it particularly well suited to OTU data. We develop a closed form expression for the Gini index, under the assumption that individual nonzero abundances follow a Pareto distribution, and we extend this expression to accommodate the additional zeros present in OTU data. We also derive the exact (non-asymptotic) distribution of the maximum likelihood estimator of the Pareto parameter b and use this to develop a confidence interval for the true Gini index. These results form a solid theoretical foundation for applying the Gini index to OTU data. In Section 4.3, we will develop a procedure that uses the Gini index to perform common/rare OTU classification.

There is only one definition for the Gini index, but there are several equivalent ways to calculate it. All are based on the Lorenz curve, which is derived from the underlying distribution of X. In economic applications, X represents the wealth or income of individuals. For OTU data, X represents the abundance of an OTU. For the current application, X designates an individual

abundance for an OTU at a site, and the sample consists of the abundances for this OTU across all sites in the data set.

Assume X is continuous with pdf f, cdf F, mean μ and support [0, ∞). The Lorenz curve consists of pairs $\left(L_x, L_y\right)$ defined by the parametric equations

$$L_x\left(w\right) = \int_0^w f\left(t\right)dt = F\left(w\right)$$

$$L_y\left(w\right) = \frac{1}{\mu}\int_0^w t \cdot f\left(t\right)dt$$

Since F is a continuous distribution, it is reasonable to assume F is one-to-one, so F^{-1} exists. With this assumption, the Lorenz curve can be written in terms of $p = L_x\left(w\right)$ as follows

$$L_y = \frac{1}{\mu}\int_0^w t \cdot f\left(t\right)dt$$

Substitute $u = F\left(t\right)$ so $F^{-1}\left(u\right) = t$ and $du = f\left(t\right)dt$.

$$L_y = \frac{1}{\mu}\int_0^{F(w)} F^{-1}\left(u\right)du$$

Let $p = F\left(w\right)$, which is the x value on the Lorenz curve. Then $L_y\left(p\right) = \frac{1}{\mu}\int_0^p F^{-1}\left(t\right)dt$ and the Lorenz curve is specified by the pairs $\left(p, L_y\left(p\right)\right)$.

The Gini coefficient, denoted by G, is defined as twice the area between the line $L_y\left(p\right) = p$ and the Lorenz curve. Note that the line is always above the curve, so absolute values are not needed to obtain the vertical distance. There are several equivalent ways to specify the area. Using the inverse of the cdf,

$$\begin{aligned} G &= 2 \times \int_0^1 \left\{p - L_y\left(p\right)\right\}dp \\ &= 2 \times \left\{\tfrac{1}{2}p^2\Big|_0^1\right\} - \frac{2}{\mu} \times \int_0^1 \left\{\int_0^p F^{-1}\left(t\right)dt\right\}dp \\ &= 1 - \frac{2}{\mu} \times \int_0^1 \left\{\int_0^p F^{-1}\left(t\right)dt\right\}dp \end{aligned}$$

If the inverse is intractable, we can work with the cdf, beginning with the parametric definition of the Lorenz curve.

$$G = 2 \times \int_0^\infty \left\{ L_x(w) - L_y(w) \right\} dL_x(w)$$
$$= 2 \times \int_0^\infty \left\{ F(w) - \frac{1}{\mu} \int_0^w t \cdot f(t)\, dt \right\} f(w)\, dw$$
$$= 2 \times \left\{ \int_0^\infty F(w) f(w)\, dw - \frac{1}{\mu} \int_0^\infty \left[\int_0^w t \cdot f(t)\, dt \right] f(w)\, dw \right\}$$

In the first integral, substitute $u = F(w)$, so $du = f(w)\, dw$. In the second integral, reverse the order of the double integral.

$$G = 2 \times \left\{ \frac{1}{2} \left[F(w) \right]^2 \Big|_0^\infty - \frac{1}{\mu} \int_0^\infty \int_t^\infty \left[f(w)\, dw \right] t \cdot f(t)\, dt \right\}$$
$$= 2 \times \frac{1}{2} - \frac{2}{\mu} \int_0^\infty \left[1 - F(t) \right] t \cdot f(t)\, dt$$
$$= 1 - \frac{2}{\mu} \int_0^\infty t \cdot f(t)\, dt + \frac{2}{\mu} \int_0^\infty F(t) \cdot t \cdot f(t)\, dt$$
$$= -1 + \frac{2}{\mu} \int_0^\infty F(t) \cdot t \cdot f(t)\, dt$$
$$= \frac{1}{\mu} \left\{ \int_0^\infty 2F(t) \cdot t \cdot f(t)\, dt - \mu \right\}$$
$$= \frac{1}{\mu} \left\{ \int_0^\infty 2F(t) \cdot t \cdot f(t)\, dt - \int_0^\infty t \cdot f(t)\, dt \right\}$$
$$= \frac{1}{\mu} \left\{ \int_0^\infty \left[2F(t) - 1 \right] t \cdot f(t)\, dt \right\}$$

This is the definition provided by Sandstrom *et al.*, (1988). Equivalent forms are given by Gastwirth (1972) and Peng (2011). These are, respectively,

$$G = \frac{1}{\mu} \int_0^\infty F(x) \left[1 - F(x) \right] dx$$
$$G = 1 - \frac{1}{\mu} \int_0^\infty \left[1 - F(x) \right]^2 dx$$

For many common distributions, the cdf is not available in a simple closed form. In this event, numeric approximations can be used to estimate the true value of the Gini coefficient. The exponential and Pareto distributions are two exceptions in that both the cdf and the Gini coefficient can be calculated directly.

Example

Assume $X \sim \exp(\beta)$, so $f(x) = \beta e^{-\beta x}$, $F(x) = 1 - e^{-\beta x}$, and $\mu = \frac{1}{\beta}$. Using Gastwirth's definition,

$$\begin{aligned} G &= \frac{1}{\mu}\int_0^\infty F(x)\left[1 - F(x)\right] dx \\ &= \beta\int_0^\infty \left(1 - e^{-\beta x}\right) e^{-\beta x} dx \\ &= \left(-e^{-\beta x} + \frac{1}{2}e^{-2\beta x}\right)\Big|_0^\infty \\ &= -(0-1) + \frac{1}{2}(0-1) \\ &= \frac{1}{2} \end{aligned}$$

Note that this does not depend on the value of β. ■

We now develop a closed form for the theoretical value for the Gini index under the Pareto distribution. We first consider the case in which there are no zeros.

Claim:

Let Y represent the nonzero abundances for an OTU (across the sites), with $Y \sim$ Pareto (1, b), and suppose that $b > 1$. Then the Gini index is $G = \frac{1}{2b-1}$.

Proof:

Y has pdf $f(y) = b \cdot y^{-b-1} I(y \geq 1)$ and cdf $F(y) = 1 - y^{-b}$, $y \geq 1$, with mean $\mu = \frac{b}{b-1}$, $b > 1$. The Gini index is

$$G_Y = \frac{1}{\mu}\int_1^\infty F(y)\left[1-F(y)\right]dy = \left(\frac{b-1}{b}\right)\int_1^\infty \left(1-y^{-b}\right)\left(y^{-b}\right)dy$$

$$= \left(\frac{b-1}{b}\right)\int_1^\infty \left(y^{-b}-y^{-2b}\right)dy$$

$$= \left(\frac{b-1}{b}\right)\left[\frac{1}{-b+1}y^{-b+1}-\frac{1}{-2b+1}y^{-2b+1}\right]_1^\infty$$

$$= \left(\frac{b-1}{b}\right)\left(\frac{0-1}{-b+1}-\frac{0-1}{-2b+1}\right) = \left(\frac{b-1}{b}\right)\left(\frac{1}{b-1}-\frac{1}{2b-1}\right)$$

$$= \left(\frac{b-1}{b}\right)\left(\frac{(2b-1)-(b-1)}{(b-1)(2b-1)}\right) = \left(\frac{b-1}{b}\right)\left(\frac{b}{(b-1)(2b-1)}\right)$$

$$= \frac{1}{2b-1}$$ ■

We now extend this to accommodate the zeros.

Claim:

Suppose X follows the zero-inflated Pareto mixture distribution with pdf

$$h(x) = \pi \cdot I(x=0) + (1-\pi)b \cdot x^{-b-1} \cdot I(x \geq 1)$$

Then the Gini index is $G = \pi + (1-\pi)\cdot\frac{1}{2b-1}$.

Proof:

X has cdf

$$H(x) = \Pr(X \leq x) = \begin{cases} 0 & x < 0 \\ \pi & 0 \leq x < 1 \\ \pi + (1-\pi)\left(1-x^{-b}\right) & x \geq 1 \end{cases}$$

and mean $E(X) = \pi \cdot 0 + (1-\pi)E(Y) = (1-\pi)\left(\frac{b}{b-1}\right)$

The Gini index is

$$
\begin{aligned}
G_X &= \frac{1}{\mu}\int_0^{\infty} H(x)\left[1-H(x)\right]dx \\
&= \frac{b-1}{b(1-\pi)}\left\{\int_0^{1} H(x)\left[1-H(x)\right]dx + \int_1^{\infty} H(x)\left[1-H(x)\right]dx\right\} \\
&= \frac{b-1}{b(1-\pi)}\left\{\pi(1-\pi)\int_0^{1} dx + \int_1^{\infty}\left[\pi+(1-\pi)\left(1-x^{-b}\right)\right]\left[1-\pi-(1-\pi)\left(1-x^{-b}\right)\right]dx\right\} \\
&= \frac{b-1}{b(1-\pi)}\left\{\pi(1-\pi) + \int_1^{\infty}\left[\pi+(1-\pi)\left(1-x^{-b}\right)\right]\left[(1-\pi)\left(1-\left(1-x^{-b}\right)\right)\right]dx\right\} \\
&= \frac{b-1}{b(1-\pi)}\left\{\pi(1-\pi) + \int_1^{\infty}\left[\pi+(1-\pi)\left(1-x^{-b}\right)\right]\left[(1-\pi)x^{-b}\right]dx\right\} \\
&= \frac{b-1}{b(1-\pi)}\left\{\pi(1-\pi) + \int_1^{\infty}\left[\pi(1-\pi)x^{-b} + (1-\pi)^2\left(1-x^{-b}\right)x^{-b}\right]dx\right\} \\
&= \frac{b-1}{b(1-\pi)}\left\{\pi(1-\pi) + \pi(1-\pi)\frac{1}{-b+1}x^{-b+1}\Big|_1^{\infty} + (1-\pi)^2\int_1^{\infty}\left(x^{-b}-x^{-2b}\right)dx\right\} \\
&= \frac{b-1}{b(1-\pi)}\left\{\pi(1-\pi) + \pi(1-\pi)\frac{1}{-b+1}(0-1) + (1-\pi)^2\left(\frac{1}{-b+1}x^{-b+1}\Big|_1^{\infty} - \frac{1}{-2b+1}x^{-2b+1}\Big|_1^{\infty}\right)\right\} \\
&= \frac{b-1}{b(1-\pi)}\left\{\pi(1-\pi) + \pi(1-\pi)\frac{1}{b-1} + (1-\pi)^2\left(\frac{1}{-b+1}(0-1) - \frac{1}{-2b+1}(0-1)\right)\right\} \\
&= \frac{b-1}{b(1-\pi)}\left\{\pi(1-\pi) + \pi(1-\pi)\frac{1}{b-1} + (1-\pi)^2\left(\frac{1}{b-1} - \frac{1}{2b-1}\right)\right\} \\
&= \frac{b-1}{b(1-\pi)}\left\{\pi(1-\pi)\left(1+\frac{1}{b-1}\right) + (1-\pi)^2\frac{2b-1-(b-1)}{(b-1)(2b-1)}\right\} \\
&= \frac{b-1}{b(1-\pi)}\left\{\pi(1-\pi)\left(\frac{b}{b-1}\right) + (1-\pi)^2\frac{b}{(b-1)(2b-1)}\right\} \\
&= \pi + (1-\pi)\frac{1}{2b-1} \\
&= \frac{2b\pi-\pi+1-\pi}{2b-1} \\
&= \frac{2\pi(b-1)+1}{2b-1}
\end{aligned}
$$

$$
G_X = \pi + (1-\pi)\cdot\frac{1}{2b-1} \qquad \blacksquare
$$

Note that the zero-inflated Gini index is a weighted average of 1 and the Gini index that excludes zeros, where the weights are the proportion of zeros and nonzeros.

We now consider the sampling distribution of the maximum likelihood estimator of the Gini index, based on a random sample of nonzero values.

Claim:

If $(x_1, x_2, \ldots x_n)$ is an *iid* sample from Pareto(1, b),

then $\hat{b} \sim$ inverted-gamma$(\alpha = n, \beta = nb)$.

Proof:

Suppose $X \sim$ Pareto(1, b), so X has pdf $f(x) = b \cdot x^{-b-1}$, for $x \geq 1$ and $b > 0$.

Then the MLE of b is $\hat{b} = \left[\frac{1}{n}\sum_{i=1}^{n} \log(x_i)\right]^{-1}$.

Define $Y = \log(X)$. Then the cdf of Y is

$$F_Y(y) = P(Y \leq y) = P(\log X \leq y) = P(X \leq e^y)$$
$$= 1 - (e^y)^{-b} = 1 - e^{-by}$$

This is the cdf of an exponential distribution with parameter $\beta = \frac{1}{b}$, which is also gamma distribution with $\alpha = 1$ and $\beta = \frac{1}{b}$. Therefore

$$\sum_{i=1}^{n} \log(x_i) = \sum_{i=1}^{n} y_i \sim \text{gamma}\left(\alpha = n, \beta = \frac{1}{b}\right)$$

and

$$\frac{1}{n}\sum_{i=1}^{n} \log(x_i) \sim \text{gamma}\left(\alpha = n, \beta = \frac{1}{nb}\right)$$

Thus $\hat{b} = \dfrac{1}{\frac{1}{n}\sum_{i=1}^{n} \log(x_i)} \sim$ inv-gamma$(\alpha = n, \beta = nb)$. ■

Note that $E\left(\hat{b}\right)=\frac{\beta}{\alpha-1}=\frac{nb}{n-1}=\left(\frac{n}{n-1}\right)b$, so $\hat{b}$ is a biased estimator of b, but $\left(\frac{n-1}{n}\right)\hat{b}$ is unbiased. Also note that $Var\left(\hat{b}\right)=\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}=\frac{n^2b^2}{(n-1)^2(n-2)}$ for $n>2$, so the variance of the unbiased estimator is $\left(\frac{n-1}{n}\right)^2 \text{Var}\left(\hat{b}\right)=\frac{b^2}{n-2}$.

Now that we know some interesting things about the distribution of $\hat{b}$, we can use this information to construct a confidence interval for the true b (and hence a confidence interval for the true G). We know that $\hat{b}\sim$ inv-gamma$(\alpha=n,\beta=nb)$ so $\frac{\hat{b}}{b}\sim$ inv-gamma$(\alpha=n,\beta=n)$ and $\frac{b}{\hat{b}}\sim$ gamma$\left(\alpha=n,\beta=\frac{1}{n}\right)$. Let L and U represent the lower and upper $\frac{\alpha}{2}$ percentiles of the gamma$\left(n,\frac{1}{n}\right)$ distribution. Then

$$\begin{aligned}1-\alpha &= P\left(L\le\frac{b}{\hat{b}}\le U\right)\\ &= P\left(2\hat{b}L-1\le 2b-1\le 2\hat{b}U-1\right)\\ &= P\left(\frac{1}{2\hat{b}U-1}\le\frac{1}{2b-1}\le\frac{1}{2\hat{b}L-1}\right)\end{aligned}$$

so a $100(1-\alpha)\%$ confidence interval for the true value of the Gini index is

$$\left[\left(2U\hat{b}-1\right)^{-1},\left(2L\hat{b}-1\right)^{-1}\right]$$

where L and U and lower and upper $\frac{\alpha}{2}$ percentiles of gamma$\left(n,\frac{1}{n}\right)$ and $n>2$ is the number of nonzero counts.

The MLE of b has nice properties, but the estimator becomes numerically unstable for very small values of b and very small sample sizes. For this reason, we rely on a more robust estimator based on the area under the Lorenz curve approximated via trapezoids. This estimator is very stable for all sample sizes, but it is also a biased estimator. The theoretical value of the Gini coefficient is based on a continuous distribution function so that the Lorenz curve for the

population is continuous. Given a random sample from the population, the Lorenz curve is estimated by a series connected line segments and the area under this 'curve' is calculated as the sum of the areas for the underlying polygons. The true (population) Lorenz curve is always convex, and using polygons to estimate the area will always over-estimate the true area. Thus the resulting estimate for the Gini coefficient is always biased downward. This can be seen in Figure 4.20.

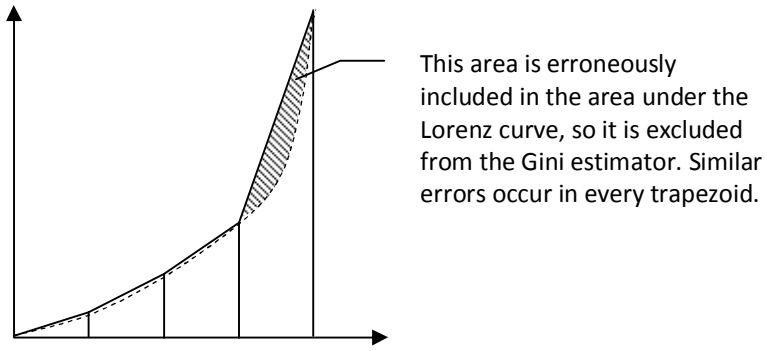


Figure 4.20: Source of Bias in the Gini Index Trapezoidal Estimator
The Gini index estimated via trapezoidal approximation excludes the shaded area and is therefore biased downward. The amount of bias is inversely proportional to the number of trapezoids, which is the number of nonzeros in the sample.

As the sample size increases, the amount of misappropriated area decreases so that the bias goes to zero. For small samples, however, the bias can be quite large and can dominant the standard error of the estimate (Deltas, 2003). The values for the true Gini coefficient range from 0 to 1. The extremes occur when either all members of the population have equal values (Gini is 1) or one member of the population has 100% of the 'wealth' and the remaining members have 0 (Gini is 0). When the Gini coefficient is estimated from a sample, the values range from 0 to $\frac{n-1}{n}$. Thus smaller samples have a reduced range for the estimated Gini coefficient, which makes comparison of different-sized samples difficult.

Deltas (2003) recommends adjusting the Gini estimates by multiplying by $\frac{n}{n-1}$. This forces the range of the Gini estimate to be [0, 1] regardless of the sample size, and increases each

estimate to alleviate the downward bias. This adjustment is needed only if the sample sizes are unequal.

This section has presented some key results about the Gini index, when the underlying distribution is Pareto with minimum value 1 and shape parameter b. These results are

- If we ignore the zeros, the true value for the Gini index is $G = \frac{1}{2b-1}$
- If we model a point mass at 0 (with proportion π), $G = \pi + (1-\pi) \cdot \frac{1}{2b-1}$.
- For a random sample of size n (excluding zeros), the exact distribution of the MLE of b is $\hat{b} \sim \text{inv-gamma}(\alpha = n, \beta = nb)$
- For a random sample of size n, a $100(1-\alpha)\%$ confidence interval for the true value of the Gini index is

  $$\left[\left(2U\hat{b} - 1 \right)^{-1}, \left(2L\hat{b} - 1 \right)^{-1} \right]$$

  where L and U and lower and upper $\frac{\alpha}{2}$ percentiles of $\text{gamma}\left(n, \frac{1}{n} \right)$
- For small values of b ($b < 1$), the theoretical value of the Gini index is undefined, but numerically stable estimates can be obtained via trapezoidal approximation.

4.3. Gini Index for Common/Rare OTU Classification

In Section 3.3.2, we introduced the Gini index as a measure for identifying rare and common OTUs. In this section, we develop a procedure for performing this classification. We are working under the premise that the values for a rare OTU follows a Pareto distribution, while the distribution for a common OTU is unspecified. We have considered (in Section 3.1.3) traditional goodness-of-fit tests, but there are two major difficulties with these tests. First, the observed values are highly skewed with sparse right tails. The chi-square goodness-of-fit test requires that the data be binned into cells, and even when the bins are defined on a logarithmic scale there are numerous cells with zero counts. Thus the asymptotic chi-square distribution may fail to hold, and the results of this test may be inaccurate. Second, the observed values typically contain many duplicate small values which results in ties among the data, thus rank-based methods such as the Kolmogorov-Smirnov test may produce unreliable results.

We propose to use the Gini index to classify OTUs as either rare or common. The range of the Gini index is [0, 1], with values near 0 indicating a more even distribution (typical of common OTUs) and value near 1 indicating a less even distribution (typical of rare OTUs). Strictly speaking, the prevalence of an OTU should also be considered when classifying OTUs, since common OTUs are generally more prevalent (occur in more sites) than rare OTUs. We do not explicitly consider the prevalence of an OTU, instead it is incorporated into the calculation of the Gini index. For each OTU, the Gini index is estimated via trapezoidal approximation (see Section 4.2) using the nonzero abundances for the OTU.

We have shown (see Section 4.2) that if the data follow a Pareto distribution with parameter $b > 1$, then the true value for the Gini index is $G = (2b-1)^{-1}$. (The true value of the Gini index is undefined if $0 < b \leq 1$, because the mean of the Pareto distribution is undefined for these values of b.) This result, however, does not translate directly to pyrosequence data sets. The derivation of the true value for the Gini index requires that the data are continuous, and the observed values in pyrosequence data sets are discrete. To derive the true value for the Gini index for discrete distributions, improper integrals are replaced by infinite sums. By the integral test, these sums are guaranteed to converge but their convergence value is unknown. Thus, for discrete counts, the true value of the Gini index is unknown. However, an estimate for the Gini index can be obtained for discrete data, but the theoretical results presented in Section 4.2 would no longer apply since the data are not continuous. The Gini index can also be estimated from the adjusted counts, which are continuous.

To obtain the sample estimate of the Gini index for each OTU, we use trapezoidal approximation. However, this estimate is known to be biased downward (Section 4.2). Furthermore, the amount of bias is proportional to the sample size (the number of nonzero counts for the OTU), so the amount of bias varies across OTUs. In addition, there is a relationship between the values of n and b. In particular, larges for one of the (n, b) pair are associated with small values of the other. For Lorena's data, this relationship is depicted in Figure 4.21. Thus both the sample size (n) and the Pareto parameter (b) affect the sampling distribution of the Gini index.

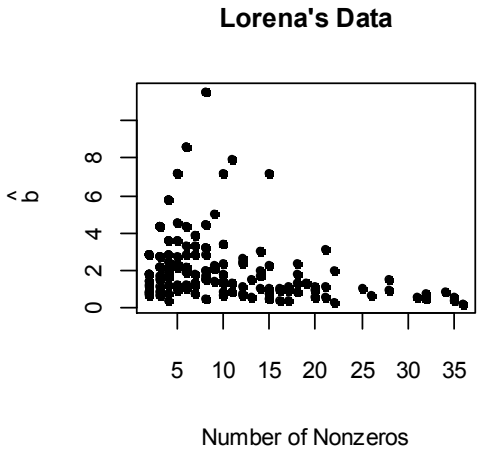


Figure 4.21: Relationship between *n* and $\hat{b}$ in Lorena's Data

For the OTUs in Lorena's data, there is an inverse relationship between the number of nonzeros for the OTU (n) and the MLE of the Pareto parameter (b).

To circumvent the uncertainties regarding the sampling distribution of the Gini values, we employ a resampling strategy. For each observed OTU, a simulated Gini distribution is derived from 1,000 parametric bootstrap samples generated from a Pareto distribution with minimum value 1 and shape parameter *b*, which is estimated from the observed sample. The size of each bootstrap sample (*n*) is equal to the observed sample size (the number of nonzero counts for the OTU). Since the Gini index is sensitive to the values of *n* and *b*, by restricting these values to the estimates obtained from the sample, we are ensuring that the simulated Gini values are comparable to the observed Gini value.

We are comparing the observed Gini value for an unknown type of OTU to a simulated distribution of Gini values generated for a rare OTU. If the OTU is common, its Gini value is likely to be smaller (closer to 0) and thus should be smaller than most of the simulated rare Gini values. We count the proportion of simulated Gini values that are smaller than the observed Gini value. We expect to see a large proportion if the unknown OTU is rare and a small proportion if the unknown OTU is common. Thus this proportion can be interpreted as a measure of the likelihood that the OTU is rare. Although this is not a true probability, for convenience we will refer to this as the probability of rare. We first apply this procedure to the original counts in Lorena's data, then we repeat the process using the adjusted counts and compare the results.

We illustrate this concept using two OTUs from Lorena's data. For each OTU, we record *n* and calculate $\hat{b}$. We generate 1,000 samples of size *n* from a Pareto distribution with shape

parameter $\hat{b}$ and calculate the Gini index for each sample. This generates one bootstrapped distribution, and we repeat this process to generate 20 bootstrapped distributions for each OTU. These are shown in Figure 4.22 and Figure 4.23. Before we assess the 'rarity' of these OTUs, notice the stability of the bootstrapped distributions. In particular, the medians and quartiles change very little within each OTU. This stability occurred for each of 218 OTUs in Lorena's data, so we feel comfortable making assessments based on a single bootstrapped distribution.

As shown in Figure 4.22 and Figure 4.23, both of the examined OTUs occur in 15 sites, but OTU 19 has total count 141 with estimated Gini value 0.42, while OTU 52 has total count 68 with estimated Gini value 0.53. On the basis of total count, we would expect OTU 19 to be 'more common' and OTU 52 to be 'more rare', but we need to consider the distribution of these counts. The bootstrap results are used to quantify these assessments. For OTU 19, the probability of rare is estimated to be slightly less than 0.25, since less than 25% of the simulated Gini values are below the observed Gini value (as indicate by the horizontal line). For OTU 52, the probability of rare is estimated at slightly above 0.5, since slightly more than one-half of the simulated Gini values are below the observed Gini value.

We apply this procedure to the original counts in Lorena's data, which contains 799 OTUs. Since the MLE of *b* involves the reciprocal of the log counts, OTUs that occur in only one site or occur only as singletons have unreliable (or undefined) estimates for *b*. These OTUs are pre-designated as rare because there is insufficient information to conclude otherwise. This removes 581 of the original 799 OTUs. The remaining 218 OTUs are classified according to the proportion of bootstrapped Gini values that are smaller than the observed Gini value for the OTU. If this proportion exceeds a pre-defined threshold, the OTU is classified as rare. Otherwise, the OTU is classified as common.

The next consideration is the choice of an appropriate threshold. We are trying to separate rare and common OTUs, and these are distinguished by fairly even counts for common OTUs and erratic counts for rare OTUs. We therefore choose a threshold that produces clear differences in the distributions of total OTU count. When viewed on a log scale, we want the total count for common OTUs to appear roughly symmetric, while the rare OTUs will remain right-skewed. The histograms in Figure 4.24 show these distributions as the threshold changes from 0.20 to 0.80. For example, when the threshold is 0.20, an OTU will be classified as rare if its estimated probability of rare is 0.20 or higher. This occurred for 192 of the 218 OTUs in

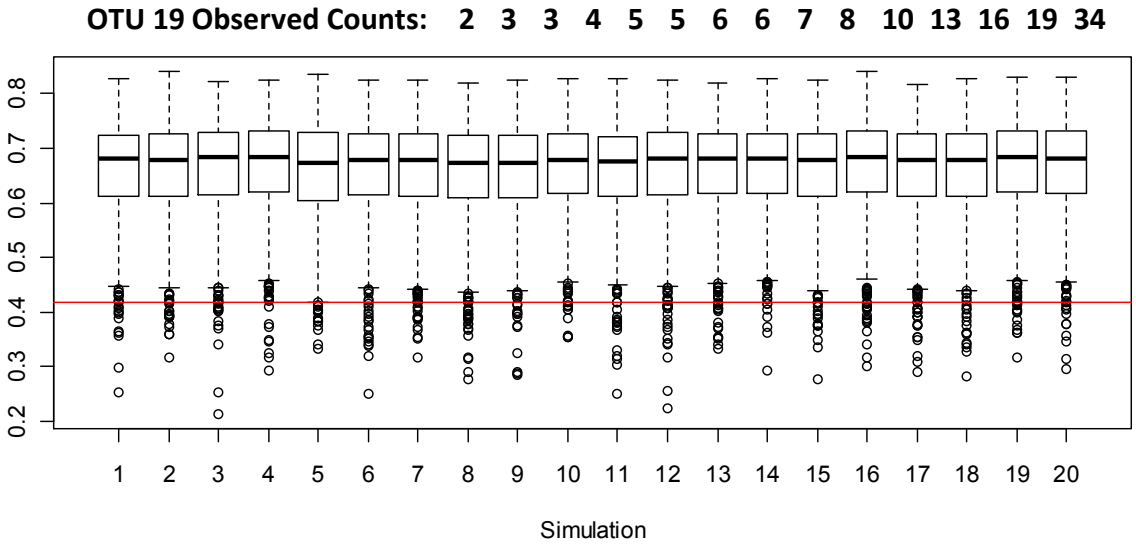


Figure 4.22: Simulated Sampling Distributions of the Gini Index for OTU 19

Each boxplot represents one simulated distribution of Gini values based on 1,000 bootstrapped samples. The horizontal line is the Gini value calculated from observed counts for this OTU. The probability that this OTU is rare is estimated to be less than 0.25, since less than 25% of the simulated distribution is below the observed value.

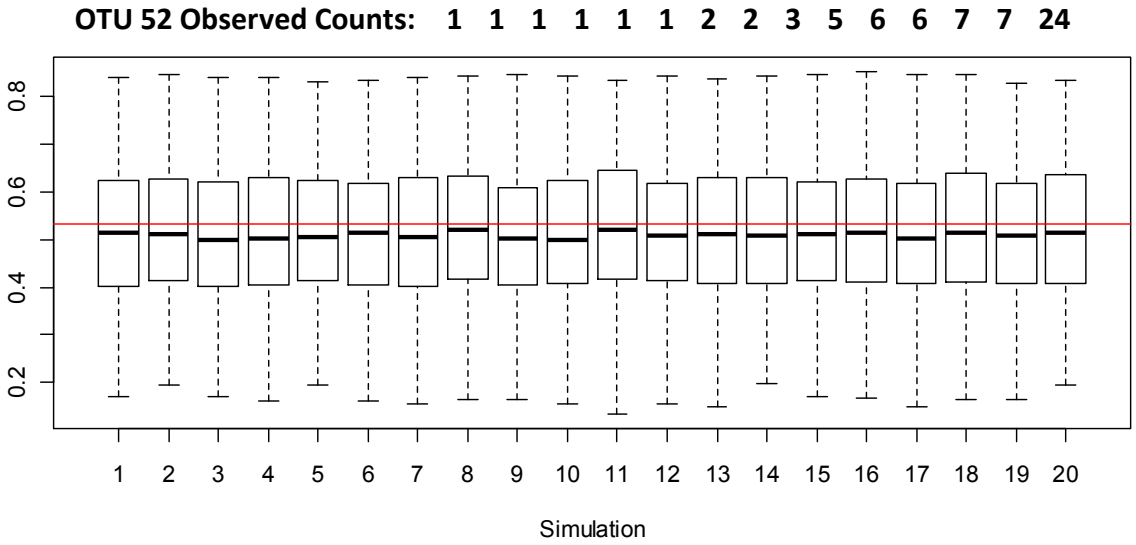


Figure 4.23: Simulated Sampling Distribution of the Gini Index for OTU 52

Each boxplot represents one simulated distribution of Gini values based on 1,000 bootstrapped samples. The horizontal line is the Gini value calculated from observed counts for this OTU. The probability that this OTU is rare is estimated to be just over 0.50, since slightly more than half of the simulated distribution is below the observed value.

Lorena's data. When the threshold is higher, some of these 'rare' OTUs will no longer be classified as rare; instead they will be classified as common. Thus as the threshold increases, more OTUs are shifted from the rare classification to the common classification. We know we have shifted too many OTUs when the common distribution becomes skewed. In Lorena's data, an appropriate threshold is somewhere between 0.40 and 0.60, so we choose the threshold 0.50.

In the bootstrapping procedure, we used the original counts to calculate the Gini index and to generate the estimate $\hat{b}$ from which the bootstrap samples were generated. Since $\hat{b}$ is unreliable or undefined for many OTUs, we were required to pre-designate many OTUs as rare. We now examine the impact of using adjusted counts to calculate the Gini index, where the counts are adjusted according to the results of the multiplicative model presented in Section 4.1.1.

In the previous bootstrapping procedure, we simply ignored the site-specific information for each of these counts, and treated all counts as equally likely. We now incorporate the results of the fitted multiplicative model by using the adjusted counts. These counts have been adjusted by the row effect for the site in which the count occurs. The reason for this modification is intuitive. All of the counts for one site are obtained from a single analyte collected at the site, and this analyte undergoes DNA extraction and amplification before being sequenced. Thus any variation the DNA extraction and amplification occurs for all counts that are observed at the site. This variation is captured in the row effects, and therefore the adjusted counts remove this variability from the observations.

Recall that Lorena's data contains one site that has a much lower total count than the remaining sites. The small site was deemed improperly amplified and was simply removed from the data set. We kept this site in the multiplicative model, and its estimated row effect is 0.5, while the row effects for the remaining sites range between 0.8 and 1.3. Since most of the row effects are close to 1, the row adjustment will not substantially alter the observed count. In contrast, the column effects vary from 0.6 to 71.9, so removing these effect from the observed counts will create very different adjusted counts. However, the column adjustment affect all counts in the same column equally. That is, there is only one effect value for each column, and every observed count in that column is adjusted (divided) by the same column effect value. The Gini index is invariant to changes in scale, so the estimate of the Gini index is unaffected by the

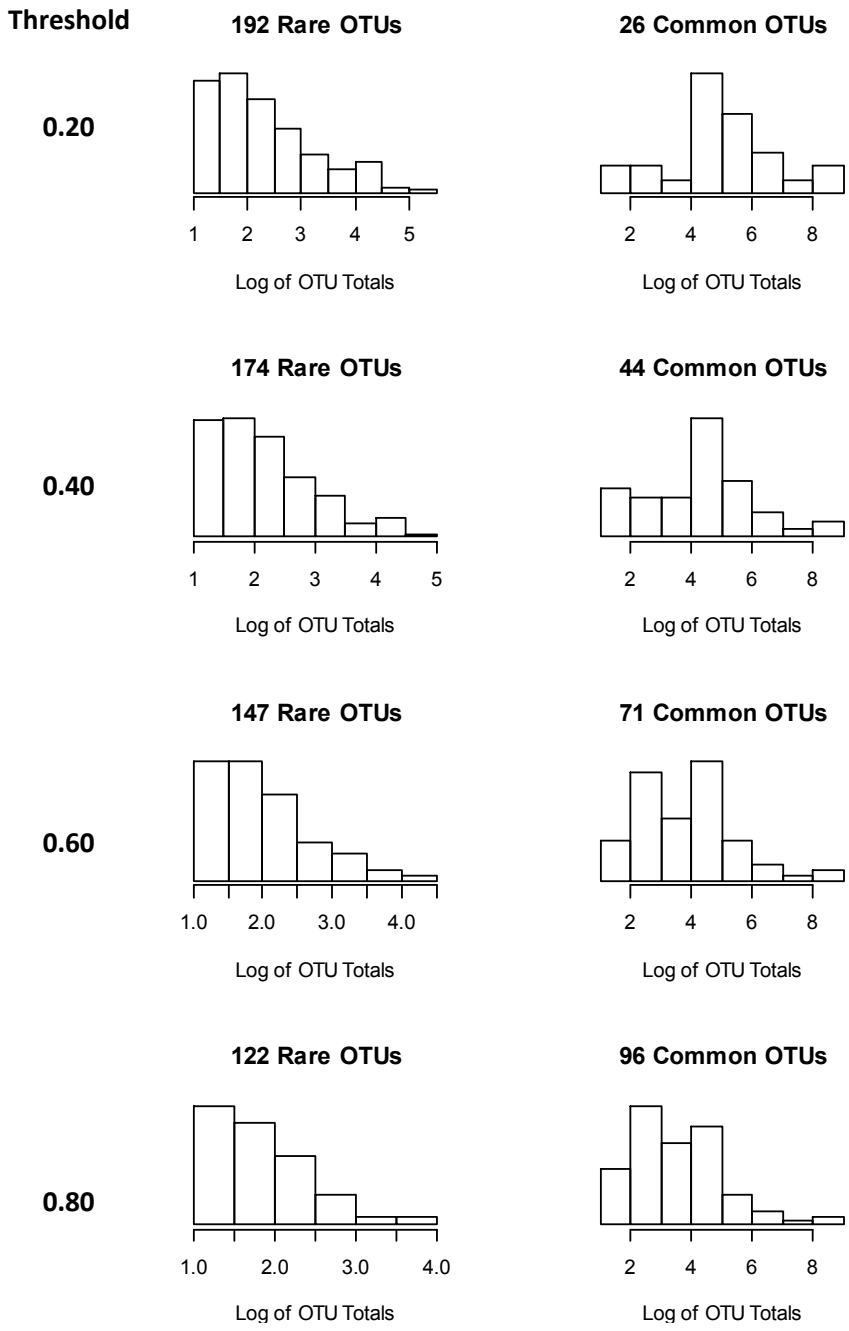


Figure 4.24: Distribution of OTU Totals as the Threshold Changes

These histograms are based on the log of total count for each OTU, using the original counts in Lorena's data. When the threshold is low, more OTUs are classified as rare. As the threshold increases, OTUs are shifted from the rare classification to the common classification. An appropriate threshold will generate a non-skewed distribution for the common OTUs. From these histograms, we can determine that an appropriate threshold is approximately 0.50.

column adjustment. Therefore only the row adjustment will affect the Gini estimate and since the row effects are not widespread, we do not expect the Gini estimate to change much when adjusted values are used. This is reflected in Figure 4.25, which compares the Gini index calculated from original counts to the index calculated from adjusted counts. The biggest differences occur when the original Gini index is 0, that is, when the original counts are all equal. After adjusting these counts via the multiplicative model, the values are no longer equal and Gini index increases as a result.

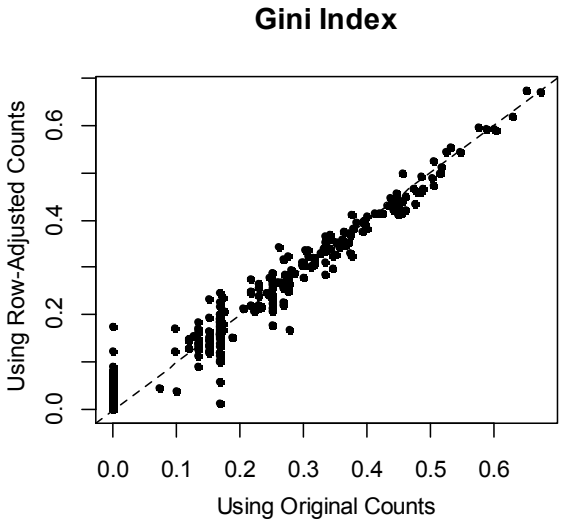


Figure 4.25: Gini Index: Original vs. Adjusted Counts

This graph compares the Gini index calculated from the original counts (x-axis) vs. adjusted counts (y-axis). The biggest change occurs for OTUs whose original counts are all equal. For these OTUs, the original Gini index is 0 because the counts are perfectly evenly distributed. When these counts are adjusted, the counts are no longer equal and the Gini index increases.

Since the values of the Gini index are not dramatically affected when the data are adjusted, we do not expect a substantial difference in the Common/Rare classification that uses this index. This is reflected in Figure 4.26, which indicates that the two methods generated the same classification in all but 23 OTUs. The circled points represent OTUs that would be classified differently depending on whether we use the original counts or the adjusted counts. Both of these methods use a threshold of 0.5, so OTUs with Pr(Rare) greater than 0.5 are classified rare and all others are classified common.

Most of the mismatched classifications in the upper left corner of Figure 4.26 correspond to OTUs that occur in a small number of sites and also have duplicate counts. When the original counts are used, the Gini index interprets the duplicate counts as arising from a equitable distribution (which is indicative of Common OTUs), and therefore erroneously classifies these OTUs as Common. When the integer counts are adjusted, the Gini index is better able to detect the rarity of the OTU. Most of the mismatched classifications in the lower right corner correspond to OTUs that occur in many sites, but also have several singletons. All but one of these points are at or near the threshold of 0.5. The exception is OTU 76, which has Pr(Rare) 0.7 using the original counts and 0.3 using the adjusted counts. This OTU occurs in 18 sites (which indicates that it is common), but it occurs as a singleton in 10 of these sites and its maximum individual count is only 4 (which indicates that it is rare). It is therefore not surprising that the two procedures would arrive at opposite classifications, since the true classification of this OTU is ambiguous.

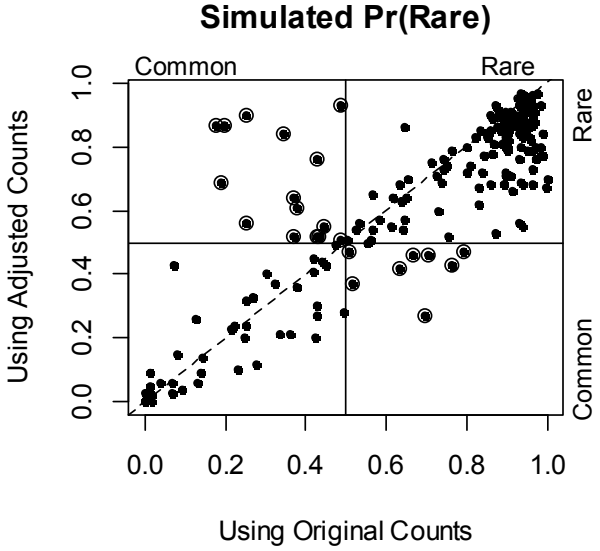


Figure 4.26: Common/Rare Classifications Using Original vs. Adjusted Counts
This graph compares the classification of each OTU, according to whether the original counts or the adjusted counts are used. Classifications were different for only 23 OTUs, shown as circled points on the graph.

It may be possible to use the estimated column effects from the multiplicative model to classify OTUs as rare or common. Each column effect is an indication of the prevalence of an OTU, both in terms of its persistence and abundance, and what the multiplicative model

perceives as excess variability may actually be a manifestation of a very common OTU. It seems logical, then, that the estimated column effects may be a suitable measure to distinguish common and rare OTUs. In our previous work, OTUs were designated as either rare or common by estimating the probability that the OTU is rare, which was based on the Gini index for the OTU. We now compare these two measures. Common OTUs should have small values from the Gini index and large column effects, while rare OTUs should have large values from the Gini index and small column effects. To orient these two measures, we compare the column effects to the complement of the probability of rare.

In Figure 4.27, the *x* axis represents the column effects (on a log scale) estimated from the multiplicative model, and the *y* axis is 1 – Pr(Rare) estimated from the Gini index. The gray circles represent the OTUs that were pre-designated as rare during the Gini procedure. These OTUs have such extremely small counts that the algorithm employed in the Gini procedure became numerically unstable for these OTUs, and they were pre-designated as rare and assigned Pr(Rare) equal to 1. The horizontal dashed line at 0.5 represents the threshold value used in the Gini procedure to designate an OTU as either rare or common. Points above this line indicate common OTUs and points below the line indicate rare OTUs.

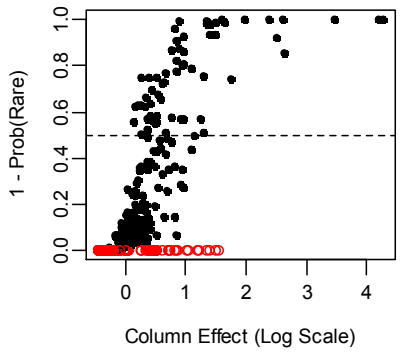


Figure 4.27: Compare Column Effects to the Probability of Rare

Each point represents one OTU from Lorena's data. Large values for the column effect (x-axis) indicates the OTU is common. The y-axis is the complement of the probability of rare, so large values also indicate the OTU is common.

The scatterplot in Figure 4.27 exhibits a positive relationship between estimates obtained in the multiplicative model (on the *x* axis) and estimates obtained via the Gini index (on the *y* axis). This seems to indicate that both measures are detecting a similar pattern in the data. This

graph also contains a tight cluster of points in the lower left corner, which corresponds to the numerous infrequent OTUs in the data. It seems, then, that either measure could be used to classify rare and common OTUs, but there are advantages and disadvantages of each measure.

One advantage of using the measure based on the Gini index is that its values are more evenly dispersed, even though they are constrained in the interval [0, 1]. The column effects have no upper bound, but they are restricted to be strictly positive. As the boxplots in Figure 4.28 illustrate, the both distributions are skewed, but the unbounded column effects have large outliers, while the measure based on the Gini index does not. One major drawback of using the Gini index as a classification measure is that it is based on the assumption that the individual counts for rare OTUs follow a Pareto distribution. While we do have evidence to support this assumption in the four data sets we are analyzing, there is no guarantee that this assumption would be satisfied by other OTU data sets. The column effects, in contrast, are completely data-driven and make no distributional assumptions about the data. It seems, then, that either of these measures could be used classify common and rare OTUs.

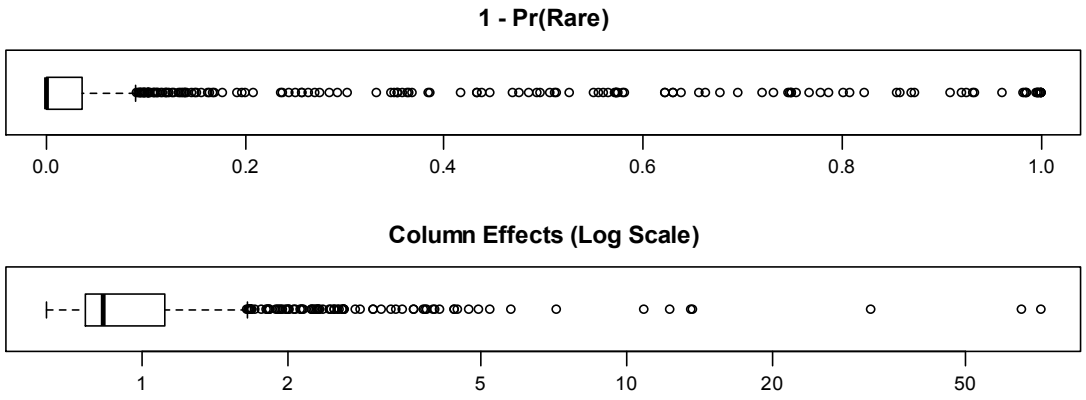


Figure 4.28: Distributions of Two Measures to Classify Common/Rare

The top boxplot is the measure based on the Gini index, and the bottom boxplot shows the column effects from the multiplicative model. Either measure could be used to classify OTUs as common or rare.

Chapter 5. Conclusion

5.1. Summary of Primary Results

With their high proportion of zeros, highly skewed nonzero values, and the potential variability in the data collection process, pyrosequence data sets must be preprocessed before they can be analyzed. For this purpose, we have proposed a multiplicative model that measures the variability in both sites and OTUs, and it generates standardized data which can be analyzed using traditional methods. The parameters of the model, which consist of an effect parameter for each site and each OTU, are estimated via a multiplicative adaption of Tukey's median polish. The residuals of the fitted model are the standardized data. The multiplicative standardization process has many advantages.

- **All of the observed data can be retained.**
  This is unlike other forms of preprocessing pyrosequence data, which require the elimination of many extremely small values and OTUs that occur in small abundances.
- **It reduces the extreme skew of the distribution.**
  This is most easily seen when both the original and standardized data are in log scale, as shown in Figure 5.1 below.
- **It highlights values that are unusual for the combination of site and OTU.**
  Relatively obscure original values can be highly unusual because they occur in an unlikely site or because they are unusually large or small for a particular OTU. The standardized values incorporate both site and OTU information to make unusual values more prominent. This was illustrated in Figure 4.14, in which an obscure original count of 47 (for OTU 31 at site 112) was standardized to 12.3. Since the standardized values are the result of a multiplicative model, a standardized value of 1 indicates a typical count, so that the standardized value of 12.3 indicates this is approximately 12 times larger than what we should expect to see for this OTU at this site.
- **The multiplicative algorithm is more efficient than traditional least squares.**
  The multiplicative model is equivalent to a log linear model, and the parameters of the multiplicative model could be estimated by traditional least squares methods. Available software for least squares estimation manipulates the design matrix of the log linear model, and these matrices can be quite large for some pyrosequence data sets. The

multiplicative adaption of Tukey's median polish requires fewer computer resources and executes in less time than traditional least squares estimation.

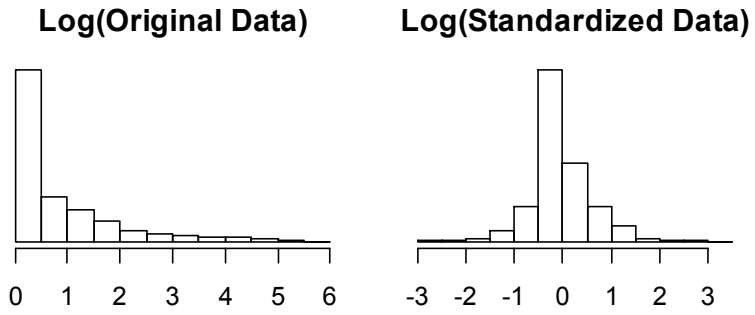


Figure 5.1: Distributions Original and Standardized Data

Both the original and standardized data are extremely skewed, but standardization reduces the degree of skewness. When the data are log-transformed, the original data are still skewed, while the standardized data are almost symmetric.

A second major contribution of this research is the development of a new procedure for classifying rare and common OTUs. This new procedure is based on the recognition that both the original data and the standardized data are highly skewed, and similar in structure to financial data such as insurance claims and the distribution of wealth in a population. Financial applications frequently use the Gini index to summarize the inequity, or lack of evenness, in a distribution. Building on the similarities between these two types of data, the Gini index is used as test statistic for performing common/rare classification. For continuous data that follows a Pareto distribution, the true value of the Gini index is known. However, the original data are discrete so the true Gini value is not known. Parametric bootstrapping yields a simulated sampling distribution of the Gini index for each OTU, from which we estimate the probability that the OTU is rare. This method appears to be far more suitable for microscopic species data than the 50% persistence threshold criterion used for macroscopic species classification, but the accuracy of the new method cannot be verified since the true classifications of the observed OTUs are unknown.

The research topics presented in this report should establish a firm foundation upon which future research can be based. To date, development of sound procedures of analysis have

been hindered by the complexities in OTU data, resulting in extensive use of standard statistical procedures with little regard for their suitability to the data. This research as well as other current and future research will likely identify new open questions to be explored. As methodologists address these questions, new tools will be added to the armamentarium of researchers producing pyrosequence data. As time progresses, these tools will be vetted for their usefulness and validity in answering questions of scientific interest, and eventually some consensus reached for which approaches seem most valid. This progression, again, will mirror the methodological development for microarray data (see Allison *et al*., 2006; Mehta *et al*., 2006). The results of the current research are seen as an early contribution to addressing the needs of researchers producing data from pyrosequencing technology.

5.2. Areas of Future Research

5.2.1. Simulate Data

A necessary component of any proposed statistical method is the ability to ascertain whether or not the method is able to uncover the 'truth' about the data generating process. For example, in Section 4.3, we developed a method to classify each OTU as either rare or common, but we could not assign any measure of confidence to the resulting classifications because we do not know which OTUs are truly rare and which ones are common. In order to make this determination, we must be able to generate simulated datasets so that we know the 'truth'.

During our investigation, we have employed two types of simulation strategies to generate portions of OTU data. These were presented in Section 4.3 and Section 4.1.4. While these strategies varied in both scope and intent, they each provide tantalizing clues as to how entire pyrosequence data sets may be simulated.

In Section 4.3, individual nonzero abundances for specific OTUs were generated by assuming the abundances followed a Pareto distribution. For each OTU that was observed in data, nonzero values were simulated specifically for this OTU, using both the sample size n and the estimated Pareto parameter b based on the observed data for that OTU. These simulations were performed under the assumption that the OTU is rare, which justifies the use of a Pareto distribution. Only nonzero values were generated, and the association between the sites and the

generated values was not modeled. That is, we generated only a vector of nonzero values for each OTU, for the purpose of estimating the sampling distribution of the Gini index for the OTU.

A different strategy was employed in Section 4.1.4. Using the results of the fitted multiplicative model, the estimated overall, row and column effects were all held fixed and the residuals were permuted. A new data set was generated by multiplying the three estimated effects and the permuted residuals. The new data set generated new estimates for the overall, row and column effects. By replicating this process we were able to simulate the sampling distributions of the estimated effects. Unlike the OTU-specific simulations in Section 4.3, the residual permutation strategy *does* associate each simulated value to an OTU and a site.

Neither of these simulation strategies explicitly model the zeros. The zeros were completely ignored in the OTU-specific simulations in Section 4.3, because the intent was to generate a distribution for the Gini index and the Gini index uses only the positive values. In the residual permutation strategy of Section 4.1.4, we begin with residuals associated with the nonzero values and permute them among the nonzero values. Thus all the observations that were originally zero will remain zero.

The ability to simulate realistic OTU data is crucial for evaluating both existing and proposed statistical methods, so this will necessarily be an area of future research. One possible approach is to extend the residual permutation strategy as applied to the multiplicative model. Instead of using residuals and effects estimated from an existing dataset, these could be generated from plausible probability models. Appropriate distributions for the row and column effects would need to be determined, and it is anticipated that the column effects will need to be generated separately for rare and common OTUs. A mixture distribution, perhaps combining two normal densities, may be suitable for modeling the asymmetric residuals, and a Bernoulli component could be used to model the zeros.

5.2.2. Measure Relationships between OTUs

In addition to enumerating OTUs that are present at any given site, researchers are often interested in identifying and quantifying the interactions and interdependencies between OTUs. For example, are there collections of OTUs that tend to occur (or not occur) together at the same sites? Or perhaps the presence (or increased abundance) of one OTU is related to a change in abundance of another OTU. While the co-occurrence patterns are straightforward to model,

other forms of relationships are difficult to identify, in part because the association may be nonlinear so that customary measures such as correlation may fail to detect it. Another obstacle in uncovering these relationships may lie in the extreme skewness of the observed data. This may be one reason the data are routinely divided into rare and common OTUs: the common OTUs occur in such large abundances that comparison to the smaller abundances of rare OTUs can be distorted. It has already been shown (in Section 4.1.5), that the standardized data reveal patterns in the data that are obscured in the original data. It seems reasonable to investigate whether the standardized data may also reveal associations between OTUs that are not evident in the original data.

5.2.3. Experimental Designs

Dr. Ari Jumpponen (personal communication, June 2012) is collecting data in a block design experiment, specifically constructed to isolate levels of variation in OTU community structure. The design involves six locations with two trees at each location, and fungal DNA is extracted from three leaves on each tree. The main objective is to compare the variability in each stratum: between leaves within the same tree; between trees in the same location; and between locations. Traditionally, block designs isolate the variability, so that the lower strata (between leaves) have less variability than the higher strata (between locations), but it is uncertain if blocking will perform its intended purpose on microbial communities. Thus the results of this experiment will provide valuable information for the design of future experiments.

In an similar biomedical study, the Human Microbiome Project recently released a report detailing a first census of the microorganisms that inhabit human bodies (The Human Microbiome Project Consortium, 2012). They found, among other things, that both the occurrence and abundance of microbes (OTUs) within one person was relatively stable over time, but that different parts of the body have very different patterns of microbial occurrence and abundance. They also report that the variation among body areas within one person can be much greater than the variation in one body area across multiple persons. While there is no direct link between microbial patterns in humans and fungal OTU patterns in a natural environment, it should not be forgotten that this is a new and emerging scientific area. Thus customary techniques, including customary experimental designs, should not be universally applied without consideration of details that are "application specific."

References

Aitchison J and JJ Egozcue (2005). Compositional Data Analysis: Where Are We and Where Should We Be Heading? *Mathematical Geology*, **37**, 829-850.

Aitchison J (2003). *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. Chapman and Hall Ltd., London, 1986. (Reprinted in 2003 with additional material by The Blackburn Press) 416 pp.

Aitchison J and M Greenacre (2002). Biplots of compositional data. *Applied Statistics*, **51**, Part 4, 375-392

Allison DB, X Cui, GP Page, and M Sabripour (2006). Microarray data analysis: From disarray to consolidation and consensus. *Nature Review Genetics*, **7**, 55–65

Anderson MJ (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology,* **26**, 32-46

Anderson MJ (2005). PERMANOVA: a FORTRAN computer program for permutational multivariate analysis of variance. Department of Statistics, University of Auckland, New Zealand. Retrieved on August 16, 2011, from http://www.stat.auckland.ac.nz/~mja/Programs.htm

Balzer S, K Malde, A Lanzén, A Sharma, and I Jonassen (2010). Characteristics of 454 pyrosequencing data – enabling realistic simulation with flowsim. *Bioinformatics*, **26**, i410-i425. DOI: 10.1093/bioinformatics/btq365

Balzer S, K Malde, and I Jonassen (2011). Systematic exploration of error sources in pyrosequencing flowgram data. *Bioinformatics*, **27**, i304-i309. DOI: 10.1093/bioinformatics/btr251

Barceló-Vidal C, JA Martin-Fernández, and V Pawlowsky-Glahn (2001). Mathematical Foundations of Compositional Data Analysis. In G. Ross (Editor) Proceedings of IAMG '01 - The sixth annual conference of the International Association for Mathematical Geology. Kansas Geological Society, Lawrence KS (CD-ROM).

Beguelin A and G Nutt (1994). Visual Parallel Programming and Determinancy - A Language Specification, an Analysis Technique, and a Programming Tool. *Journal of Parallel and Distributed Computing*, **22**, 235-250. DOI: 10.1006/jpdc.1994.1084

Bellemain E, T Carlsen, C Brochmann, E Coissac, P Taberlet, and H Kauserud (2010). ITS as an environmental DNA barcode for fungi: an in silico approach reveals potential PCR biases. *BMC Microbiology*, **10**:189

Billheimer D, P Guttorp, and WF Fagan (2001). Statistical Interpretation of Species Composition. *Journal of the American Statistical Association Applications and Case Studies*, **96**, 1205-1214

Burnham KP and PS Overton (1978). Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika*, **65**, 927-936

Cha SH (2007) Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions. *International Journal of Mathematical Models and Methods in Applied Sciences* **1**, 300-307

Chao A (1984). Non-parametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics*, **11**, 265-270

Chayes F (1960). On correlation between variables of constant sum. *Journal of Geophysical Research*, **65**, 4185-4193

Clarke KR and RM Warwick (2001) *Change in marine communities: an approach to statistical analysis and interpretation*, 2nd ed. Plymouth Marine Laboratory, UK:PRIMER-E Ltd.

Damgaard C and J Weiner (2000). Describing inequality in plant size or fecundity. *Ecology*, **81**, 1139-1142

Daunis-I-Estadella J, C Barceló-Vidal, and A Buccianti (2006). *Exploratory compositional data analysis*. From: Buccianti, A., Mateu-Figueras, G, & Pawlowsky-Glahn, V. (eds) *Compositional Data Analysis in the Geosciences: From Theory to Practice.* Geological Society, London, Special Publications, **264**, 161-174

Deltas G (2003)., The Small Sample Bias of the Gini Coefficient: Results and Implications for Empirical Research. *The Review of Economics and Statistics*, **85**, 226-234

Dickie IA (2010). Insidious effects of sequencing errors on perceived diversity in molecular surveys. *New Phytologist*, **188**, 916-918

Dixon P (2003). VEGAN, A Package of R Functions for Community Ecology. *Journal of Vegetation Science*, **14**, 927-930

Efron B and RJ Tibshirani (1994). *An Introduction to the Bootstrap.* Chapman & Hall/CRC, Boca Raton FL.

Efron B (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association*, **99,** 96-104

Faith DP, PR Minchin, and L Belbin (1987). Compositional dissimilarity as a robust measure of ecological distance. *Vegetatio* **69**, 57-68

Fischer C, A Flohre, LW Clement, P Batáry, WW Weisser, T Tscharntke, and T Carsten (2011). Mixed effects of landscape structure and farming practice on bird diversity. *Agriculture, Ecosystems and Environment*, **141**, 119-125

Fisher RA, AS Corbet and CB Williams (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology* **12**, 42-598

Fry JM, TRL Fry, and KR McLaren (2000). Compositional data analysis and zeros in micro data. *Applied Economics*, **32**, 953-959

Fujiyoshi M, S Yoshitake, K Watanabe, K Murota, Y Tsuchiya, M Uchida and T Nakatsubo (2011). Successional changes in ectomycorrhizal fungi associated with the polar willow Salix polaris in a deglaciated area in the High Arctic, Svalbard. *Polar Biology*, **34**, 667-673

Gadbury GL, PP Grier, J Edwards, T Kayo, TA Prolla, R Weindruch, PA Permana, J Mountz, DB Allison (2004). Power and Sample Size Estimation in High Dimensional Biology. *Statistical Methods in Medical Research*, **13**, 325-338.

Gao Q and ZL Yang (2010). Ectomycorrhizal fungi associated with two species of *Kobresia* in an alpine meadow in the eastern Himalaya. *Mycorrhiza*, **20**, 281-287

Gastwirth J, R Modarres and E Bura (2005). The use of the Lorenz curve, Gini index and related measures of relative inequality and uniformity in securities law. *Metron - International Journal of Statistics*, **63**, 451-469

Gastwirth J (1975). Statistical Measures of Earnings Differentials. *The American Statistician*, **29**, 32-35

Gastwirth JL (1972). The Estimation of the Lorenz Curve and Gini Index. *The Review of Economics and Statistics*, **54**, 306-316

Geml J, GA Laursen, IC Herriott, JM McFarland, MG Booth, N Lennon, HC Nusbaum and DL Taylor (2010). Phylogenetic and ecological analyses of soil and sporocarp DNA sequences reveal high diversity and strong habitat partitioning in the boreal ectomycorrhizal genus *Russula* (Russulales; Basidiomycota). *New Phytologist*, **187**, 494-507

Gilles A, E Meglécz, N Pech, S Ferreira, T Malausa, and JF Martin (2011). Accuracy and quality assessment of 454 GL-FLX Titanium pyrosequencing. *BMC Genomics*, **12**:245

Gotelli NJ and RK Colwell (2001). Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, **4**, 379-391

Greenacre M (2009). Distributional Equivalence and Subcompositional Coherence in the Analysis of Compositional Data, Contingency Tables and Ratio-Scale Measurements. *Journal of Classification*, **26**, 29-54

Huang XQ and A Madan (1999). CAP3: a DNA sequence assembly program. *Genome Research,* **9**, 868-877

The Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, **486**, 207-214

Huse SM, JA Huber, HG Morrison, ML Sogin and DM Welch (2007). Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology*, **8**:R143. DOI: 10.1186/gb-2007-8-7-r143

Irizarry RA, B Hobbs, F Collin, et al. (2003a). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249-264. DOI: 10.1093/biostatistics/4.2.249

Irizarry RA, BM Bolstad, F Collin, LM Cope, B Hobbs and TP Speed (2003b). Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research*, **31**, **e15**. DOI: 10.1093/nar/gng015.

Ishak HD, R Plowes, R Sen, K Kellner, E Meyer, DA Estrada, SE Dowd and UG Mueller (2011). Bacterial Diversity in *Solenopsis invicta* and *Solenopsis geminata* Ant Colonies Characterized by 16S amplicon 454 Pyrosequencing. *Microbial Ecology*, **61**, 821-831

Jumpponen A and KL Jones (2010a). Seasonally dynamic fungal communities in the *Quercus macrocarpa* phyllosphere differ between urban and nonurban environments. *New Phytologist* **186**, 496-513. DOI: 10.1111/j.1469-8137.2010.03197.x

Jumpponen A, KL Jones and J Blair (2010b). Vertical distribution of fungal communities in tallgrass prairie soil. *Mycologia* **102**, 1027-1041

Jumpponen A, K Keating, G Gadbury, KL Jones and JD Mattox (2010). Multi-element fingerprinting and high throughput sequencing identify multiple elements that affect fungal communities in *Quercus macrocarpa* foliage. *Plant Signaling & Behavior*, **5**, 1-5

Jumpponen A and KL Jones (2009). Massively parallel 454 sequencing indicates hyperdiverse fungal communities in temperate *Quercus macrocarpa* phyllosphere. *New Phytologist* **184,** 438-448. DOI: 10.1111/j.1469-8137.2009.02990.x

Kafadar K and T Phang (2003). Transformation, background estimation, and process effects in the statistical analysis of microarrays. *Computational Statistics and Data Analysis*, **44**, 313-338

Kao WC and YS Song (2011). naiveBayesCall: An Efficient Model-Based Base-Calling Algorithm for High-Throughput Sequencing. *Journal of Computational Biology*, **18**, 365-377. DOI: 10.1089/cmb.2010.0247

Kao WC, S Kristian, and YS Song (2009). BayesCall: A model-based base-calling algorithm for high-throughput short-read sequencing. *Genome Research*, **19**, 1884-1895. DOI: 10.1101/gr.095299.109

Krebs CJ (1999). *Ecological Methodology*. Addison-Wesley Educational Publishers, Inc.

Kunin V and P Hugenholtz (2010a). PyroTagger: A fast, accurate pipeline for analysis of rRNA amplicon pyrosequence data. *The Open Journal, Article 1*, 1-8 Retrieved on August 13, 2011 from http://www.theopenjournal.org/toj_articles/1

Kunin V, A Engelbrektson, H Ochman, and P Hugenholtz (2010b). Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environmental Biology*, **12**, 118-123

Ledergerber C and C Dessimoz (2010). Base-calling for next-generation sequencing platforms. Briefings in Bioinformatics, Advance Access published January 18, 2011. DOI: 10.1093/bib/bbq077

Lee C, RG Klopp, R Weindruch and TA Prolla (1999). Gene expression profile of aging and its retardation by caloric restriction. *Science* **285**, 1390-1393.

Legendre P and L Legendre (1998) *Numerical Ecology*, Second English Edition. Elsevier Science B.V.

Lozupone CA, M Hamady, ST Kelley and R Knight (2007). Quantitative and Qualitative β Diversity Measures Lead to Different Insights into Factors That Structure Microbial Communities. *Applied and Environmental Biology*, **73**, 1576-1585

MacArthur R (1957), On the relative abundance of bird species. *Proceedings of the National Academy of Sciences,* **43**, 293-295.

Magurran AE and PA Henderson (2003). Explaining the excess of rare species in natural species abundance distributions. *Nature*, **422**, 714-716

Magurran AE (2004) *Measuring Biological Diversity*, Blackwell Publishing

Margulies M, M Egholm, WE Altman, S Attiya, JS Bader, LA Bemben, J Berka, MS Braverman, YJ Chen, Z Chen, SB Dewell, L Du, JM Fierro, XV Gomes, BC Godwin, W He, S Helgesen, CH Ho, GP Irzyk, SC Jando, MLI Alenquer, TP Jarvie, KB Jirage, JB Kim, JR Knight, JR Lanza, JH Leamon, SM Lefkowitz, M Lei, J Li, KL Lohman, H Lu, VB Makhijani, KE McDade, MP McKenna, EW Myers, E Nickerson, JR Nobile, R Plant, BP Puc, MT Ronan, GT Roth, GJ Sarkis, JF Simons, JW Simpson, M Srinivasan, KR Tartaro, A Tomasz, KA Vogt, GA Volkmer, SH Wang, Y Wang, MP Weiner, P Yu, RF Begley, and JM Rothberg (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376-380

Martín-Fernández JA and S Thió-Henestrosa (2006). Rounded zeros: some practical aspects for compositional data. From A. Buccianti, G. Mateu-Figueras and V. Pawlowsky-Glahn (eds), *Compositional Data Analysis in the Geosciences: From Theory to Practice*. Geological Society, London, Special Publications, **264**, 191-201

Martin-Fernández JA, C Barceló-Vidal, and V Pawlowsky-Glahn (2003). Dealing with Zeros and Missing Values in Compositional Data Sets Using Nonparametric Imputation. *Mathematical Geology*, **35**, 253-278

Martin-Fernández JA, C Barceló-Vidal, and V Pawlowsky-Glahn (2000). Zero replacement in compositional data set. In H. Kiers, J. Rasson, P. Groenen, and M. Shader, eds., *Studies in classification, data analysis, and knowledge organization*, Proceedings of the 7th Conference of the International Federation of Classification Societies (IFCS '2000), University of Namur, Namur: Springer-Verlag, Berlin, 155-160

McGill BJ, RS Etienne, JS Gray, D Alonso, MJ Anderson, HK Benecha, M Dornelas, BJ Enquist, JL Green, F He, AH Hurlbert, AE Magurran, PA Marquet, BA Maurer, A Ostling, CU Soykan, KI Ugland and EP White (2007). Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecology Letters*, **10**, 995-1015

Mehta TS, SO Zakharkin, GL Gadbury, and DB Allison (2006). Epistemological issues in omics and high-dimensional biology: Give the people what they want. *Physiological Genomics,* **28**, 24-32

Mouillot D and A LePretre (2000). Introduction of Relative Abundance Distribution (RAD) Indices, Estimated from the Rank-Frequency Diagrams (RFD), to Assess Changes in Community Diversity. *Environmental Monitoring and Assessment,* **63**, 279-295

Nyrén P (2007). The History of Pyrosequencing. From *Methods of Molecular Biology*, *vol. 373: Pyrosequencing Protocols*. Edited by S. Marsh, Humana Press Inc., Totowa NJ, 1-14

Palarea-Albaladejo J, JA Martín-Fernández, and J Gómez-García (2007). A Parametric Approach for Dealing with Compositional Rounded Zeros. *Mathematical Geology*, **39**, 625-645

Pawlowsky-Glahn V and JJ Egozcue (2006). Compositional data and their analysis: an introduction. From A. Buccianti, G. Mateu-Figueras and V. Pawlowsky-Glahn (eds), *Compositional Data Analysis in the Geosciences: From Theory to Practice*. Geological Society, London, Special Publications, **264**, 1-10

Pawlowsky-Glahn V and JJ Egozcue (2002)., BLU Estimators and Compositional Data. *Mathematical Geology*, **34**, 259-274

Pearson K (1897). Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London*, **60**, 489-502

Peng L (2011). Empirical Likelihood Methods for the Gini Index. *Australian and New Zealand Journal of Statistics*, **53**, 131-139

Quinlan AR, DA Stewart, and MP Stromberg (2008). Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nature Methods*, **5**, 179-181. DOI: 10.1038/nmeth.1172

Quince C, A Lanzén, TP Curtis, RJ Davenport, N Hall, IM Head, LF Read, and WT Sloan (2009). Accurate determination of microbial diversity from 454 pyrosequencing data. *Nature Methods*, **6**, 639-694. DOI: 10.1038/nmeth.1361

Quince C, A Lanzen, RJ Davenport, and PJ Turnbaugh (2011). Removing Noise From Pyrosequenced Amplicons. *BMC Bioinformatics,* **12**, Article Number 38, DOI: 10.1186/1471-2105-12-38

R Development Core Team (2009). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org

Reeder J and R Knight (2010). Rapid denoising of pyrosequencing amplicon data: exploiting the rank-abundance distribution. *Nature Methods*, **7**, 668-669

Roesch LFW, RR Fulthorpe, A Riva, G Casella, AKM Hadwin, AD Kent, SH Daroub, FAO Camargo, WG Farmerie and EW Triplett (2007). Pyrosequencing enumerates and contrasts soil microbial diversity. *The ISME Journal*, **1**, 283-290.

Ronaghi M, S Karamohamed, B Pettersson, M Uhlén, and P Nyrén (1996). Real-time DNA sequencing using detection of pyrophosphate release. *Analytical Biochemistry*, **242**, 84-89. DOI: 10.1006/abio.1996.0432

Ronaghi M, M Uhlén, and P Nyrén (1998). A sequencing method based on real-time pyrophosphate detection. *Science*, **281**, 363-365

Sandstrom A, JH Wretman, and B Walden (1988). Variance Estimators of the Gini Coefficient - Probability Sampling. *Journal of Business & Economic Statistics*, **6**, 113-119

Selby B (1965). The index of dispersion as a test statistic. *Biometrika*, **52**, 627-629.

Simon RM and K Dobbin (2003). Experimental design of DNA microarray experiments. *Bio Techniques*, **34**, 16-21

Smith EP and G van Belle (1984). Nonparametric estimation of species richness. *Biometrics*, **40**, 119-129

Storey JD (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society - Series B*, **64**, 479-498

Tauber F (1999). Spurious Clusters in Granulometric Data Caused by Logratio Transformation. *Mathematical Geology*, **31**, 491-504

Unterseher M, A Jumpponen, M Opik, L Tedersoo, M Moora, CF Dormann and M Schnittler (2011). Species abundance distributions and richness estimations in fungal metagenomics - lessons learned from community ecology. *Molecular Ecology*, **20**, 273-285.

van der Gast CJ, AW Walker, FA Stressmann, GB Rogers, P Scott, TW Daniels, MP Carroll, J Parkhill and KD Bruce (2011). Partitioning core and satellite taxa from with cystic fibrosis lung bacterial communities. *The ISME Journal*, **5**, 780-791.

Van Diepen LTA, EA Lilleskov and KS Pregitzer (2011). Simulated nitrogen deposition affects community structure of arbuscular mycorrhizal fungi in northern hardwood forests. *Molecular Ecology*, **20**, 799-811

Weisstein EW. "Dirichlet Integrals." Retrieved April 5, 2011 from *Mathworld* --A Wolfram Web Resource. http://mathworld.wolfram.com/DirichletIntegrals.html

Wharton DI and FKC Hui (2011). The arcsine is asinine: the analysis of proportions in ecology. *Ecology,* **92**, 3-10

Williamson M and KJ Gaston (2005). The lognormal distribution is not an appropriate null hypothesis for the species-abundance distribution. *Journal of Animal Ecology*, **74**, 409-422

Williamson M (2010). Why do species abundance distributions of individuals and of biomass behave differently under sampling? *Oikos*, **119**, 1697-1699

Zhang MQ (1998). Statistical features of human exons and their flanking regions. *Human Molecular Genetics*, **7**, 919-932. DOI: 10.1093/hmg/7.5.919

Zhang Z, S Schwartz, L Wagner and W Miller (2000). A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology*, **7**, 203-214

Zimmerman GM, H Goetz, PW Mielke, Jr. (1985). Use of an Improved Statistical Method for Group Comparisons to Study Effects of Prairie Fire. *Ecology*, **66**, 606-611

Appendix A. Glossary

Amalgamation

The act of combining two or more parts of a composition. For example, if we take a 4-part composition (x_1, x_2, x_3, x_4) and amalgamate parts 3 and 4, the result is the 3-part composition $(y_1, y_2, y_3) = (x_1, x_2, x_3 + x_4)$.

Base pairs

Each strand of double-stranded DNA contains a sequence of nucleotides. The two strands are joined into a double helix by chemical bonds between two nucleotides, one nucleotide on each strand. The pair of nucleotides that are bonded together are called base pairs.

Closure

The transformation in which each element of a positive-valued vector is divided by the sum of the elements. The image is a compositional vector. The closure operation is denoted by ?.

Composition

For an integer $D \geq 2$, a D-part composition is defined to be $\boldsymbol{x} = (x_1, x_2, \ldots, x_D)$, where $x_i \geq 0$ and $\sum_{i=1}^{D} x_i = 1$. For logratio analysis, the x_i are required to be strictly positive.

Diversity

A combination of species richness and species richness that can be used to measure the health of an ecological community or to measure differences between communities.

DNA

Deoxyribonucleic acid, consisting of two strands of nucleotides coiled together to form a double helix.

Evenness

The equity of species abundances in a community.

Genome

The complete DNA sequence of an organism.

OTU

Operational Taxonomic Unit. In the current research, OTU is a surrogate for species.

PCR amplification

Polymerase Chain Reaction: one method for duplicating (amplifying) sections of DNA

Perturbation

A vector operation on the simplex, equivalent to vector addition in real space. The perturbation operation is denoted by $\oplus$. For a composition $\boldsymbol{x}$ and any vector $\boldsymbol{u}$ that has nonnegative elements,

$$\boldsymbol{u} \oplus \boldsymbol{x} = \mathcal{C}\left(u_1 x_1, u_2 x_2, \ldots, u_D x_D\right) = \frac{\left(u_1 x_1, u_2 x_2, \ldots, u_D x_D\right)}{u_1 x_1 + u_2 x_2 + \ldots + u_D x_D}$$

Powering

A vector operation on the simplex, equivalent to scalar multiplication in real space. The powering operation is denoted by $\odot$. For a scalar $\alpha \in \Re^1$ and compositional vector $\boldsymbol{x}$,

$$\alpha \odot \boldsymbol{x} = \mathcal{C}\left(x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha\right) = \frac{\left(x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha\right)}{x_1^\alpha + x_2^\alpha + \ldots + x_D^\alpha}$$

Simplex

The set of all D-part compositions, denoted ∇^{D-1}.

$$\nabla^{D-1} = \left\{\left(x_1, x_2, \ldots, x_D\right) \mid x_1 \geq 0, x_2 \geq 0, \ldots, x_D \geq 0; x_1 + x_2 + \ldots + x_D = 1\right\}.$$

Singletons

Singletons occur when an OTU is observed exactly once at a site. This is recorded as a '1' in the OTU data matrix.

Richness

The number of species present at a site.

Subcomposition

A composition containing a subset of parts. For example, if the original composition contains four parts $\left(x_1, x_2, x_3, x_4\right)$, a two-part subcomposition is $\mathcal{C}\left(x_1, x_2\right) = \left(\frac{x_1}{x_1 + x_2}, \frac{x_2}{x_1 + x_2}\right)$

Appendix B. Log Series Distribution and Fisher's α

The logarithmic series distribution is a discrete distribution defined by the probability mass function

$$P(T=k)=\frac{a\theta^k}{k},\quad 0<\theta<1,\quad k=1,2,\ldots,\quad \text{where } a=\frac{-1}{\ln(1-\theta)}.$$

Since the value for a is completely defined in terms of θ, this distribution has only one parameter. This distribution is right-skewed and is more strongly skewed as θ approaches 1, as shown in Figure B.1.

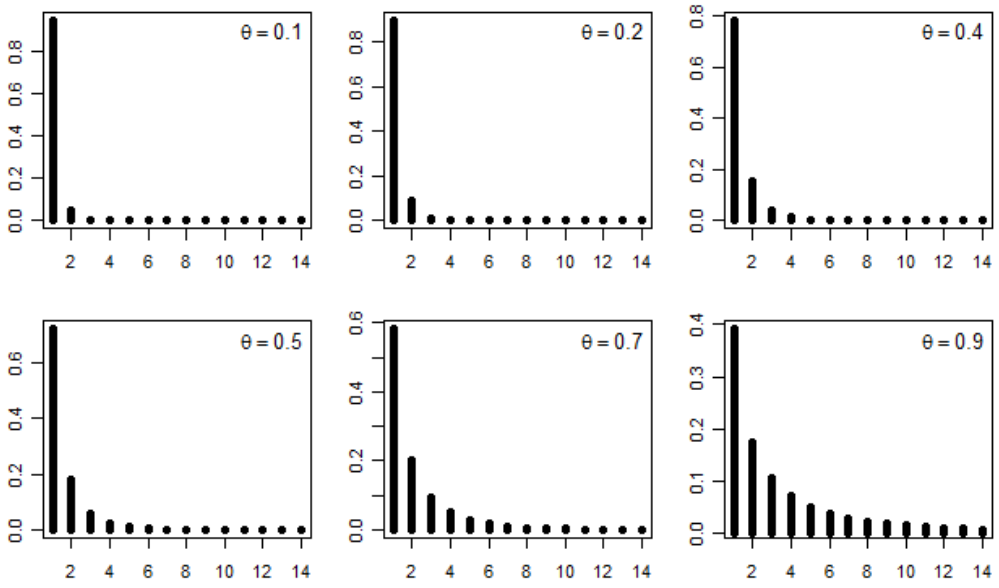


Figure B.1: Log Series Distribution

To use this distribution for species abundances, let $X_i, i=1,2,\ldots S,$ be the total abundances (across all sites) for OTU i and let $N=\sum_{i=1}^{S}X_i$ be the total abundance for all OTUs.

To find the maximum likelihood estimator of θ, we treat the elements of the vector **X** as independent observations and construct the log likelihood function

$$\ell(\theta\,|\,\boldsymbol{x})=\sum_{i=1}^{S}\left\{\ln\left(\frac{-1}{\ln(1-\theta)}\right)+x_i\ln\theta-\ln x_i\right\}$$

With repeated use of the chain rule, the derivative is

$$\begin{aligned}\frac{\partial}{\partial\theta}\ell(\theta) &= \sum_{i=1}^{S}\left\{\frac{1}{-1\cdot\left[\ln(1-\theta)\right]^{-1}}\cdot\left[\ln(1-\theta)\right]^{-2}\cdot\frac{-1}{1-\theta}+\frac{x_i}{\theta}\right\}\\ &= \sum_{i=1}^{S}\frac{1}{(1-\theta)\ln(1-\theta)}+\sum_{i=1}^{S}\frac{x_i}{\theta}\\ &= \frac{S}{(1-\theta)\ln(1-\theta)}+\frac{N}{\theta}\end{aligned}$$

The MLE $\hat{\theta}$ is the solution to $\frac{\partial\ell}{\partial\theta}=0$, which is the solution to

$$\frac{N}{\hat{\theta}}=\frac{-S}{(1-\hat{\theta})\cdot\ln(1-\hat{\theta})} \quad \text{or} \quad \frac{S}{N}=\frac{(\hat{\theta}-1)\cdot\ln(1-\hat{\theta})}{\hat{\theta}}.$$

Note that this implies $\ln(1-\hat{\theta})=\frac{S}{N}\cdot\frac{\hat{\theta}}{(\hat{\theta}-1)}$.

By the invariance property of MLEs,

$$\hat{a}=\frac{-1}{\ln(1-\hat{\theta})}=\frac{-1}{\frac{S}{N}\cdot\frac{\hat{\theta}}{(\hat{\theta}-1)}}=\frac{N(1-\hat{\theta})}{S\cdot\hat{\theta}}.$$

This is the MLE for *a* in the log series *distribution*, which has a one-to-one correspondence to the log series used in ecological studies. The log series is defined by

$$\alpha\theta,\frac{\alpha\theta^2}{2},\frac{\alpha\theta^3}{3},\ldots,\frac{\alpha\theta^k}{k},\ldots$$

where the k^{th} term in the series is the predicted number of OTUs that have abundance *k*. Thus the relationship between the series and the distribution is

$$P(\text{an OTU has count } k)=\frac{\text{number of OTUs with count } k}{\text{number of OTUs}}$$

$$\frac{a\theta^k}{k}=\frac{1}{S}\cdot\frac{\alpha\theta^k}{k}$$

$$S\cdot a=\alpha$$

The value of Fisher's α is the MLE of α, $\hat{\alpha}=S\cdot\hat{a}=S\cdot\frac{N}{S}\cdot\frac{1-\hat{\theta}}{\hat{\theta}}=N\cdot\frac{1-\hat{\theta}}{\hat{\theta}}$.

Appendix C. Proof of Result 3.4

Result 3.4:

Let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ be the order statistics of a random sample of size *n* from a Pareto (1, *b*) distribution with pdf $f(x) = b \cdot x^{-b-1}, \quad x \geq 1, \quad b > 0$. Then the ratio $\frac{X_{(k+1)}}{X_{(k)}}$ follows a Pareto distribution with parameters 1 and $b(n-k)$.

Proof:

The pdf of X is $f(x) = b \cdot x^{-b-1}, \quad x \geq 1, \quad b > 0$, and the cdf is $F(x) = \begin{cases} 0 & \text{if } x < 1 \\ 1 - x^{-b} & \text{if } x \geq 1 \end{cases}$.

For $i < j$ and $x_i < x_j$, the joint pdf of the order statistics $X_{(i)}$ and $X_{(j)}$ is

$$f_{i,j}(x_i, x_j) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!} f(x_i) f(x_j) \left[F(x_i)\right]^{i-1} \left[F(x_j) - F(x_i)\right]^{j-i-1} \left[1 - F(x_j)\right]^{n-j}$$

For consecutive order statistics $X_{(k)}$ and $X_{(k+1)}$, this simplifies to

$$\begin{aligned} f_{k,k+1}(x_k, x_{k+1}) &= \frac{n!}{(k-1)!(n-k-1)!} f(x_k) f(x_{k+1}) \left[F(x_k)\right]^{k-1} \left[1 - F(x_{k+1})\right]^{n-k-1} \\ &= \frac{n!}{(k-1)!(n-k-1)!} \left[b \cdot x_k^{-b-1}\right] \left[b \cdot x_{k+1}^{-b-1}\right] \left[1 - x_k^{-b}\right]^{k-1} \left[1 - \left(1 - x_{k+1}^{-b}\right)\right]^{n-k-1} \\ &= \frac{n! b^2}{(k-1)!(n-k-1)!} \cdot \frac{1}{x_k^{b+1}} \cdot \left[1 - x_k^{-b}\right]^{k-1} \cdot \frac{1}{x_{k+1}^{b+1}} \cdot \left[x_{k+1}^{-b}\right]^{n-k-1} \end{aligned}$$

Define the transformation $U = X_{(k)}$ and $V = \frac{X_{(k+1)}}{X_{(k)}}$, so the inverse transformation is

$X_{(k)} = U$ and $X_{(k+1)} = UV$, with Jacobian $J = \begin{vmatrix} 1 & 0 \\ v & u \end{vmatrix} = u$.

Then the joint pdf of U and V is

$$
\begin{aligned}
f_{U,V}(u,v) &= f_{k,k+1}(u,uv)\cdot u \\
&= \frac{n!\,b^2}{(k-1)!(n-k-1)!}\cdot\frac{1}{u^{b+1}}\left(1-u^{-b}\right)^{k-1}\frac{1}{(uv)^{b+1}}\left[(uv)^{-b}\right]^{n-k-1}\cdot u \\
&= \frac{n!\,b^2}{(k-1)!(n-k-1)!}\cdot\frac{1}{v^{b+1+bn-bk-b}}\cdot\left(1-u^{-b}\right)^{k-1}\cdot\frac{1}{u^{2b+2+bn-bk-b-1}} \\
&= \frac{n!\,b^2}{(k-1)!(n-k-1)!}\cdot\frac{1}{v^{b(n-k)+1}}\cdot\left(1-u^{-b}\right)^{k-1}\cdot\frac{1}{u^{b+1}}\cdot\frac{1}{u^{b(n-k)}} \\
&= \frac{n!\,b^2\,v^{-b(n-k)-1}}{(k-1)!(n-k-1)!}\left(u^{-b}\right)^{n-k}\left(1-u^{-b}\right)^{k-1}u^{-b-1}
\end{aligned}
$$

We want the pdf of V, so integrate out u.

$$
f_V(v) = \frac{n!\,b^2\,v^{-b(n-k)-1}}{(k-1)!(n-k-1)!}\int_1^{\infty}\left(u^{-b}\right)^{n-k}\left(1-u^{-b}\right)^{k-1}u^{-b-1}du
$$

Substitute $t = u^{-b}$, so $dt = -b\,u^{-b-1}$. As $u \to \infty, t \to 0$ (because $b > 0$), and as $u \to 1, t \to 1$.

$$
\begin{aligned}
f_V(v) &= \frac{n!\,b^2\,v^{-b(n-k)-1}}{(k-1)!(n-k-1)!}\left(\frac{1}{-b}\right)\int_1^0 (t)^{n-k}\left(1-t\right)^{k-1}dt \\
&= \frac{n!\,b\,v^{-b(n-k)-1}}{(k-1)!(n-k-1)!}\int_0^1 (t)^{n-k}\left(1-t\right)^{k-1}dt
\end{aligned}
$$

The integral is a beta function.

$$
\begin{aligned}
f_V(v) &= \frac{n!\,b\,v^{-b(n-k)-1}}{(k-1)!(n-k-1)!}\cdot\frac{\Gamma(n-k+1)\Gamma(k)}{\Gamma(n+1)} \\
&= \frac{n!\,b\,v^{-b(n-k)-1}}{(k-1)!(n-k-1)!}\cdot\frac{(n-k)!(k-1)!}{n!} \\
&= b(n-k)v^{-b(n-k)-1}
\end{aligned}
$$

Therefore, $V = \dfrac{X_{(k+1)}}{X_{(k)}}$ follows a Pareto$\left(1, b(n-k)\right)$ distribution. ■

This pdf applies to the ratio of any two consecutive order statistics from a Pareto distribution. For the purpose of identifying potential large outliers, we are interested in the specific ratio $V = \dfrac{X_{(n)}}{X_{(n-1)}}$. For this ratio, $k = n-1$ (or $n-k=1$), and its pdf is $f_V(v) = b\,v^{-b-1}$, which is precisely the pdf of the original distribution.

Appendix D. Review of Compositional Data Analysis

Compositional data are defined to be multivariate observations whose elements are nonnegative and sum to one. From a mathematical perspective, compositional data are derived from a larger class of multivariate data known as directional data, in which each element in the vector is required to be nonnegative, but the elements do not necessarily sum to one. For OTU data, the vector of observed abundances for a particular site is a directional vector. When the vector of OTU abundances for a site is divided by the site total, the result is a vector of sample proportions, which is a compositional vector.

To facilitate comparison of sites that have different site totals some researchers analyze proportions rather than raw abundance values. Shannon's index, for example, is based on sample (site) proportions. These proportional (compositional) vectors are constrained so that the elements sum to one. If the statistical analysis is conducted on the proportions instead of the raw abundances, the constrained nature of these vectors may need to be considered. This is accomplished with compositional data analysis.

Much work has been done with this research project to determine the characteristics of OTU data and the suitability of compositional analysis for such data. A conclusion reached and presented in Section 3.2 is that, for a variety of reasons, little is to be gained from compositional analysis of OTU data, and the potential cost of doing so can be added theoretical and computational complexity and increased difficulty in interpreting results. In order to make this argument and for completeness, this appendix reviews some background of compositional analysis.

The sum-to-unity constraint for compositional data presents numerous challenges for analysis. Simple concepts that are usually taken for granted with real-valued multivariate data, such as vector addition and scalar multiplication, can generate highly unexpected and sometimes nonsensical results when applied to compositional data. For example, taking the difference of two compositional vectors will result in a vector that is not a composition because its components will no longer sum to one and because some components could be negative. These difficulties are compounded when we attempt to calculate more complicated measures, such as similarity and variability.

The transformation known as closure maps directional (count) vectors to compositional (proportion) vectors; this transformation simply re-scales each element in the directional vector by dividing each element by the sum of the element values. The elements in a compositional vector are also called parts, or components. A directional vector with S elements can be visualized as geometric ray (*i.e.* a half-line) from the origin into the positive orthant of $\Re^S$, and the act of closure projects the directional vector onto a restricted hyperplane in $\Re^{S-1}$. Thus this transformation is sometimes called the linear transformation. Other transformations have been suggested (Barceló-Vidal *et al.*, 2001), including a spherical transformation and a hyperbolic transformation, which project a ray from the positive orthant to the unit hypersphere and hyperbolic surface, respectively. The spherical transformation divides directional vector by its Euclidean (L_2) norm and hyperbolic transformation divides a directional vector by the product of its component parts. The hyperbolic transformation receives further attention in compositional data analysis as it is related to a centering transformation. (Aitchison, 2003, p.79)

When the closure transformation is applied to a directional vector with S elements, the result is a S-part composition: $x = (x_1, x_2, \ldots x_S)$, where the elements are nonnegative and sum to one. The set of all S-part compositions occupy a subspace of $\Re^{S-1}$. The reduction of dimension occurs as a direct result of the sum-to-one constraint, since $x_S = 1 - \sum_{i=1}^{S-1} x_i$. The remaining S-1 parts of the composition are also constrained by the fact that their sum cannot exceed 1. Thus the space occupied by all S-part compositions is a subspace of $\Re^{S-1}$. This subspace is called the simplex, denoted ∇^{S-1}. Since it is a constrained subspace (more specifically, a truncated hyperplane), the geometry of the simplex is unlike customary Euclidean geometry . When S = 3, the simplex can be represented geometrically by a triangular surface bounded by the points (0,0,1), (0,1,0) and (1,0,0). Data points in this space are typically represented in ternary diagrams, as shown in Figure D.1.

Extension to higher dimensions is achieved by either creating a grid of ternary diagrams, similar to scatterplot matrices, or through the use of biplots. In the Euclidean sense, a biplot represents a projection of the sample points onto the plane created by the first two principal components. This has been adapted by Aitchison and Greenacre (2002) for the geometry of simplex.

(a) The three-part simplex $\mathbb{S}^2$ as a subspace of $\mathbb{R}^3$

(b) Ternary diagram for three-part composition (x_1, x_2, x_3)

Figure D.1: Visual Representations of the Three-Part Simplex

In addition to computational challenges, much care must be taken when modeling and interpreting compositional data. The sum-to-unity constraint forces dependence among the parts of a compositional vector. Since a positive change in any one part of the composition forces a negative change in at least one other part of the composition, it would seem that the correlation between any two parts must necessarily be negative. This is not always the case, however. For example, a shortage of a particular food supply can cause the abundance of some OTUs to decay at the same rate, and thus intuitively these OTUs should be positively correlated. In addition, the usual methods for estimating correlations can produce arbitrary results when applied to compositional data. (Chayes, 1960)

In an early study, Karl Pearson (1897) warned of the dangers of "spurious correlations" that can occur between ratios of random variables. For example, X/Z and Y/Z can exhibit strong correlation even when X, Y and Z are mutually independent. In the case of pyrosequence data, we expect even stronger spurious correlations since the numerators are not independent of the common denominator (the sum). Despite Pearson's warnings, much of the work with compositional data completely disregards the constrained nature of the data. In the 1980's, John Aitchison postulated that compositional data provide information solely on the relative, and not absolute, magnitudes of the components and therefore statistical analysis must be based on ratios of components within the compositional data vector. Taking the logarithm of these ratios transforms the constrained compositional vectors into unconstrained real-valued vectors, so that conventional multivariate techniques can be applied to the transformed vectors and the results translated back into the compositional framework. This approach is called logratio analysis (Aitchison, 2003).

Although mathematically rigorous, this approach has not been universally accepted. A thorough literature review has revealed many applications within the geological sciences, but none in community ecology. There is also some dissention among geologists. Opposition has primarily focused on the difficulty in interpreting logratio results and consolidating these results with pre-existing work. Interested readers can follow the discourse in the Letters to the Editor of *Mathematical Geology* from 1988 to 2002.

D.1. Criteria for Reasonable Statistical Approaches

Any reasonable statistical analysis must generate results that are reproducible and consistent with results obtained when other valid approaches are used on the same data. For compositional data analysis, the three main principles are scale invariance, subcompositional coherence, and permutation invariance.

Scale invariance is simply a recognition that compositional data provide information only about the relative values between components, so that ratios of components are the relevant values to examine. This implies that any function f of the data must also be scale invariant, that is, for any compositional vector $\boldsymbol{x}$ and positive scalar α, $f(\alpha\boldsymbol{x}) = f(\boldsymbol{x})$.

The concept of subcompositional coherence involves working with subsets of components within a composition. For example, suppose a compositional data set contains five components, and one researcher analyzes all five components while a second researcher analyzes only the first three components. Any relationships detected in the second analysis should also be detected in the first analysis. In other words, the presence of additional components in the first analysis should not affect the relationships between the common components that are present in both analyses.

The requirement of subcompositional coherence also extends to amalgamations, in which some components are combined (added) so that the length of the compositional vector is reduced while keeping the sum-to-one constraint intact. As an illustration, suppose that analysis of a five-part compositional data set indicates strong correlation between parts 1 and 5 and between parts 2 and 5. If parts 1 and 2 are combined into a single part and the data are re-analyzed as a four-part composition, any reasonable statistical approach must be able to detect a strong correlation between the new part and the original part 5.

In addition to scale invariance and subcompositional coherence, any sensible statistical methodology must also be permutation invariant. This means that the results should be unaffected if the parts are the composition are simply rearranged (permuted). Prior to Aitchison's logratio approach, statistical techniques applied to compositional data routinely failed to achieve scale invariance, subcompositional coherence and/or permutation invariance. The result was much confusion in scientific communities, since two researchers could reach opposite conclusions when using the same data.

D.2. Logratio Transformations and Zeros

The most direct way to analyze compositional data vectors is to first transform them into real space, perform the desired multivariate analysis on the unconstrained vectors, and then transform the results back to the simplex. An acceptable transformation must map ∇^{S-1} onto $\Re^{S-1}$ and it must be one-to-one so that the results can be mapped from $\Re^{S-1}$ back to ∇^{S-1}. Typically, the transformed vectors are assumed to follow a multivariate normal distribution. For $x \in \nabla^{S-1}$ and $y \in \Re^{S-1}$, the three transformations that appear most often in the literature are the additive logratio transformation (alr), the centered logratio transformation (clr) and the isometric logratio transformation (ilr). These are defined by

$$\boldsymbol{y} = \text{alr}(\boldsymbol{x}) = \log\left(\frac{x_1}{x_S}, \frac{x_2}{x_S}, \ldots, \frac{x_{S-1}}{x_S}\right) \text{ and } \boldsymbol{y} = \text{clr}(\boldsymbol{x}) = \log\left(\frac{\boldsymbol{x}}{g(\boldsymbol{x})}\right),$$

where $g(\boldsymbol{x}) = \left(\prod_{i=1}^{S} x_i\right)^{1/S}$ is the geometric mean of x. The isometric logratio transformation generates vectors whose elements are coordinates with respect to an orthonormal basis for the simplex. The choice of the set of basis vectors dictates the precise form of the transformation.

In order to perform logratio analysis, it is necessary that all the elements of every compositional vector to be strictly positive. One mechanism for eliminating zeros is to simply amalgamate (combine) elements. This is not feasible for a data set that contains many zeros because it may require combining parts that are of central importance to the research objectives. Another strategy is to replace each zero with a small positive number prior to logratio transformation. Replacement strategies and their impact on the resulting analysis is currently an active area of research. See, for example, Fry, Fry and McLaren (2000), Palarea-Albaladejo, *et.al.* (2007), Martin-Fernandez, *et.al.* (2000, 2003, 2006), and Tauber (1999).

D.3. The Simplex as a Vector Space

The logratio transformations are well-suited for mapping the data vectors but they cannot be applied to model parameters. For example, means and standard errors of model parameters based on the simplex geometry cannot be interpreted in terms of untransformed proportions. Thus research objectives defined in terms of proportions are answered in terms of logratios. This is an unfortunate situation, so statisticians and mathematicians are currently exploring methods that exploit the structure of compositional data in order to provide interpretable answers.

When analyzing compositional data, the unique geometry of the simplex cannot be ignored. Simple measures, such as "average" can have many different meanings in the simplex. In addition, vector operations in the simplex are unlike vector operations in Euclidean space. In the simplex, the two main binary operations are *perturbations* and *powering*. These correspond to vector addition and scalar multiplication, respectively, in Euclidean space.

A perturbation models change in a composition. For example, consider a composition consisting of the proportion of three species in a habitat. Suppose that the initial composition is (0.4, 0.1, 0.5), then the habitat experiences a disturbance so that its composition changes. Further suppose that the first species is reduced by 50%, while the remaining species are reduced by 20% each. Then the result is (0.5*0.4, 0.8*0.1, 0.8*0.5) = (0.2, 0.08, 0.4), but this vector needs to be closed (so that it sums to one). The closure operation is denoted by $\mathcal{C}$. The composition after disturbance is

$$\mathcal{C}(0.5*0.4, 0.8*0.1, 0.8*0.5) = \frac{(0.2, 0.08, 0.4)}{0.2+0.08+0.4} \approx (0.294, 0.118, 0.588).$$

Note that the first proportion is reduced from 0.4 to 0.294, but the second two proportions actually increase as a result of the disturbance. This is counter-intuitive, since the actual abundances of these species are presumed to decrease. In this example, however, the first species is reduced at a larger rate so that, after the disturbance, the *relative* amounts of second two species are larger. This example illustrates one difficulty in measuring differences between compositional vectors.

In general, the perturbation operation is defined by $\boldsymbol{u} \oplus \boldsymbol{x} = \mathcal{C}(u_1 x_1, u_2 x_2, \ldots u_S x_S)$, where $\boldsymbol{x}$ is a compositional vector and the perturbing vector $\boldsymbol{u}$ contains nonnegative entries. When modeling a transition, the perturbing vector is not required to be a composition; in fact, one or more of the entries may be greater than 1. This would model a process in which a species

becomes more abundant. The powering operation in the simplex is the equivalent of scalar multiplication in Euclidean real space. For any scalar α and any composition $\boldsymbol{x}$, the powering operation is defined by $\alpha \odot \boldsymbol{x} = \mathcal{C}(x_1^\alpha, x_2^\alpha, \ldots x_S^\alpha)$

With perturbation as vector addition and powering as scalar multiplication, Billheimer, Guttorp and Fagan (2001) propose a 'linear' model involving these two operations. Their model is intuitively appealing because the parameters are directly interpretable within the framework of compositional processes. In the same article, they show that the simplex is both a vector space and a Hilbert space, with an inner product based on logratios. This provides the necessary structure from which we can define a distance between two compositions, and also allows the use of probability measures directly on the simplex.

D.4. Distance, Center and Variability

The inner product of two compositional vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined as

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_a = [\mathrm{clr}(\boldsymbol{x})]' [\mathrm{clr}(\boldsymbol{y})] = [\mathrm{alr}(\boldsymbol{x})]' \left(\mathbf{I}_{S-1} - \tfrac{1}{S}\mathbf{J}_{S-1}\right)[\mathrm{alr}(\boldsymbol{y})]$$

where **I** is the identity and **J** is a square matrix of 1's. The simplicial distance between $\boldsymbol{x}$ and $\boldsymbol{y}$ is $d_a(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\langle z, z \rangle_a}$, where $z = \boldsymbol{x} \oplus (-1 \odot \boldsymbol{y})$. (The subscript *a* denotes Aitchison's distance.) Note the similarity between the simplicial distance and ordinary Euclidean distance. In the simplex, perturbation is vector addition and powering is scalar multiplication, so $\boldsymbol{x} \oplus (-1 \odot \boldsymbol{y})$ in the simplex is equivalent to $\tilde{\boldsymbol{x}} - \tilde{\boldsymbol{y}}$ in Euclidean space. In both spaces, distance is measured as a sum of squared differences. Unlike their Euclidean counterparts, distances in the simplex are difficult to assess by visual inspection, especially when the points are near the boundary of the simplex. An example given by Billheimer, *et al.* (2001) illustrates this point. Given the three 3-part compositions $\boldsymbol{x}$ = (1/3, 1/3, 1/3), $\boldsymbol{y}$ = (0.1, 0.1, 0.8) and $\boldsymbol{z}$ = (0.01, 0.01, 0.98), the simplicial distances are $d_a(\boldsymbol{x}, \boldsymbol{y})$ = 1.698 and $d_a(\boldsymbol{x}, \boldsymbol{z})$ = 2.046. These three points are plotted in the ternary diagram in Figure D.2. By visual inspection, we see that the Euclidean distance between $\boldsymbol{y}$ and $\boldsymbol{z}$ is smaller than the distance between $\boldsymbol{x}$ and $\boldsymbol{y}$. In simplicial geometry, the distance between $\boldsymbol{y}$ and $\boldsymbol{z}$ is greater. This occurs because the constraints on the simplex cause distance to be visually distorted near the edges.

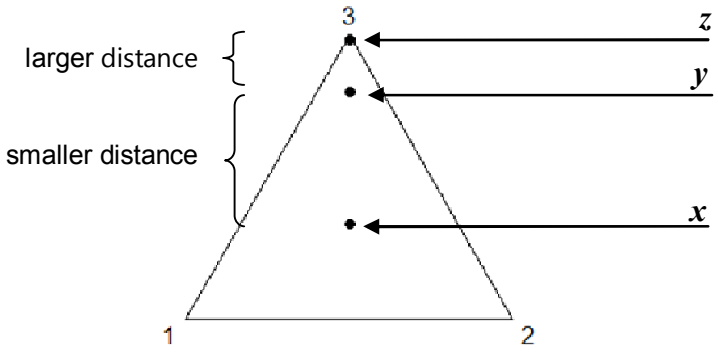


Figure D.2: Visually Deceptive Distances in the Simplex

For a random composition $\mathbf{X} = [X_1, X_2, \ldots, X_D]$, a measure of center is defined to be the point $\mathbf{a}$ in the simplex that minimizes $E[d_a(\mathbf{X}, \mathbf{a})^2]$, the mean value of the squared simplicial distance. Note that this definition parallels the definition of the expected value of a random variable in Euclidean space. This center is $\mathbf{a} = \text{cen}[\mathbf{X}] = \mathcal{C}\left[\exp\left(\text{E}[\log \boldsymbol{X}]\right)\right]$ which is also called the geometric mean (Aitchison and Egozcue, 2005). A natural estimator is $\hat{\mathbf{a}}$, the closed vector of geometric means, where the geometric mean for each component is taken across all observed vectors. This estimator is the best linear unbiased estimator (Pawlowsky-Glahn and Egozcue, 2002) in terms of simplicial geometry. In spite of this, $\hat{\mathbf{a}}$ would be impractical if the observed data contains any zeros, since this would cause the estimated center for the affected component to be 0, no matter how many positive entries may be present.

There are several approaches to measuring the variability of a random composition. Aitchison (2003) defines three measures of variation for a S-part composition:

- variation matrix: $\mathbf{T} = \left[\tau_{ij}\right] = \left[\text{var}\left(\log\left(\frac{X_i}{X_j}\right)\right)\right]$
- logratio covariance matrix: $\Sigma = \left[\sigma_{ij}\right] = \left[\text{cov}\left\{\log\left(\frac{X_i}{X_S}\right), \log\left(\frac{X_j}{X_S}\right)\right\}\right]$
- centered logratio covariance matrix: $\Gamma = \left[\gamma_{ij}\right] = \left[\text{cov}\left\{\log\left(\frac{X_i}{g(X)}\right), \log\left(\frac{X_j}{g(X)}\right)\right\}\right]$,

  where $g(X)$ is the geometric mean

Note that □ is the covariance matrix of □ = alr(□) and □ is the covariance matrix of □ = clr(□), but □ is not a covariance matrix. Pawlowsky-Glahn and Egozcue (2002) use the variation matrix □ to define the metric variance, a scalar value that represents the overall dispersion, defined by

$$\text{Mvar}(\boldsymbol{X}) = E\left[d_a^2(\boldsymbol{X},\xi)\right] = \frac{1}{S-1}\sum_{i<j}\text{var}\left(\log\frac{X_i}{X_j}\right) = \frac{1}{S-1}\sum_{i<j}\tau_{ij}$$

In order to obtain reproducible results, the use of logratio analysis is considered necessary whenever the observed multivariate observations are constrained to have unit sum. However, there are unique features of OTU data that may render logratio analysis unnecessary. Practical issues for applying logratio analysis to OTU data are discussed in Section 3.2.

D.5. Derivation of Result 3.2

Result 3.2:

Assume $X_j \sim gamma(\alpha_j, \beta)$, $j = 1, 2, \ldots, S$, with X_j and $X_{j'}$ independent for $j \neq j'$ and define $P_j = \dfrac{X_j}{\sum_{i=1}^{S} X_i}$.

Then $\text{corr}\left(P_i, P_j\right) = \dfrac{-\sqrt{\alpha_i \alpha_j}}{\sqrt{\left(\alpha_+ - \alpha_i\right)\left(\alpha_+ - \alpha_j\right)}}$, where $\alpha_+ = \sum_{k=1}^{S}\alpha_k$

Derivation:

Assume $X_j \overset{ind}{\sim} gamma(\alpha_j, \beta)$, $j = 1, 2, \ldots S$. Note that the scale parameter does not depend on j.

Define $T_{(ij)} = \sum_{k\neq i, k\neq j}^{S} X_k = T - X_i - X_j$. For simplicity of notation, let $\alpha_+ = \sum_{j=1}^{S}\alpha_j$ and

$\alpha_{ij} = \sum_{k\neq i, k\neq j}^{S}\alpha_k = \alpha_+ - \alpha_i - \alpha_j$. Then $T_{(ij)} \sim gamma(\alpha_{ij}, \beta)$, and $T_{(ij)}$, X_i, and X_j are mutually independent. The joint pdf is

$$f_{X_i, X_j, T_{(ij)}}(x_i, x_j, t) = \frac{x_i^{\alpha_i - 1} x_j^{\alpha_j - 1} t^{\alpha_{ij} - 1}}{\Gamma(\alpha_i)\Gamma(\alpha_j)\Gamma(\alpha_{ij})\beta^{\alpha_+}} \exp\left[-\frac{1}{\beta}\left(x_i + x_j + t\right)\right], \; x_i > 0, x_j > 0, t > 0$$

We will use this density to derive the joint distribution of (P_i, P_j). Define the transformation

$$U = \frac{X_i}{X_i + X_j + T_{(ij)}} \qquad V = \frac{X_j}{X_i + X_j + T_{(ij)}} \qquad Q = X_i + X_j + T_{(ij)}$$

with inverse transformation

$$X_i = UQ \qquad X_j = VQ \qquad T_{(ij)} = Q - UQ - VQ$$

and Jacobian

$$J = \begin{vmatrix} \frac{\partial x_i}{\partial u} & \frac{\partial x_i}{\partial v} & \frac{\partial x_i}{\partial q} \\ \frac{\partial x_j}{\partial u} & \frac{\partial x_j}{\partial v} & \frac{\partial x_j}{\partial q} \\ \frac{\partial T_{(ij)}}{\partial u} & \frac{\partial T_{(ij)}}{\partial v} & \frac{\partial T_{(ij)}}{\partial q} \end{vmatrix} = \begin{vmatrix} Q & 0 & U \\ 0 & Q & V \\ -Q & -Q & 1-U-V \end{vmatrix} = Q\left[Q(1-U-V)+QV\right]+UQ^2 = Q^2$$

For $u > 0$, $v > 0$, $u + v < 1$ and $q > 0$, the joint pdf of (U, V, Q) is

$$f_{U,V,Q}(u,v,q) = f_{X_i,X_j,T_{(ij)}}\left(uq, vq, q(1-u-v)\right)\cdot q^2 .$$

Let $\kappa = \left[\Gamma(\alpha_i)\Gamma(\alpha_j)\Gamma(\alpha_{ij})\beta^{\alpha_+}\right]^{-1}$. Then

$$\begin{aligned} & f_{U,V,Q}(u,v,q) \\ & = \kappa\cdot(uq)^{\alpha_i-1}(vq)^{\alpha_j-1}\left[q(1-u-v)\right]^{\alpha_{ij}-1} q^2 \exp\left[-\frac{1}{\beta}\left(uq+vq+q(1-u-v)\right)\right] \\ & = \kappa\cdot u^{\alpha_i-1}v^{\alpha_j-1}q^{\alpha_i+\alpha_j-2}q^{\alpha_{ij}-1}q^2(1-u-v)^{\alpha_{ij}-1}\exp\left(-\tfrac{q}{\beta}\right) \\ & = \kappa\cdot u^{\alpha_i-1}v^{\alpha_j-1}(1-u-v)^{\alpha_{ij}-1}q^{\alpha_+-1}\exp\left(-\tfrac{q}{\beta}\right) \end{aligned}$$

Note that U is really P_i and V is P_j, so to get the joint distribution of (P_i, P_j), we need to integrate out q.

$$\begin{aligned} f_{U,V}(u,v) &= \kappa\cdot u^{\alpha_i-1}v^{\alpha_j-1}(1-u-v)^{\alpha_{ij}-1}\int_0^\infty q^{\alpha_+-1}\exp\left(-\tfrac{q}{\beta}\right)dq \\ &= \kappa\cdot u^{\alpha_i-1}v^{\alpha_j-1}(1-u-v)^{\alpha_{ij}-1}\beta^{\alpha_+}\int_0^\infty\left(\tfrac{1}{\beta}\right)^{\alpha_+-1}q^{\alpha_+-1}\exp\left(-\tfrac{q}{\beta}\right)\left(\tfrac{1}{\beta}\right)dq \\ &= \kappa\cdot u^{\alpha_i-1}v^{\alpha_j-1}(1-u-v)^{\alpha_{ij}-1}\beta^{\alpha_+}\Gamma(\alpha_+) \end{aligned}$$

$$= \frac{\beta^{\alpha_+}\Gamma(\alpha_+)}{\Gamma(\alpha_i)\Gamma(\alpha_j)\Gamma(\alpha_{ij})\beta^{\alpha_+}} u^{\alpha_i-1}v^{\alpha_j-1}(1-u-v)^{\alpha_{ij}-1}$$

So the joint pdf of (P_i, P_j) is

$$f_{P_i,P_j}(p_i,p_j) = \frac{\Gamma(\alpha_+)}{\Gamma(\alpha_i)\Gamma(\alpha_j)\Gamma(\alpha_{ij})} p_i^{\alpha_i-1}p_j^{\alpha_j-1}(1-p_i-p_j)^{\alpha_{ij}-1}, \text{ for } p_i>0,\ p_j>0,\ p_i+p_j\le 1$$

In order to calculate $\operatorname{cov}(P_i,P_j)$, we will need to use a Dirichlet type 1 integral (Weisstein, 2011). This integral is defined by

$$\int\int\cdots\int_{\nabla} g(t_1+t_2+\cdots+t_n)t_1^{\alpha_1-1}t_2^{\alpha_2-1}\cdots t_n^{\alpha_n-1}dt_1dt_2\cdots dt_n = \frac{\prod_{i=1}^{n}\Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{n}\alpha_i\right)}\int_0^1 g(\tau)\,\tau^{\Sigma\alpha-1}\,d\tau$$

where $\nabla = \left\{(t_1,t_2,\ldots,t_n)\,\middle|\, t_i>0, \sum t_i<1\right\}$

From the definition of covariance $\operatorname{cov}(P_i,P_j) = E(P_iP_j) - E(P_i)E(P_j)$, we have

$$E(P_iP_j) = \frac{\Gamma(\alpha_+)}{\Gamma(\alpha_i)\Gamma(\alpha_j)\Gamma(\alpha_{ij})}\iint_{\nabla}(1-p_i-p_j)^{\alpha_{ij}-1}\,p_i^{\alpha_i}p_j^{\alpha_j}\,dp_idp_j$$

Let $g(p_i+p_j) = \left[1-(p_i+p_j)\right]^{\alpha_{ij}-1}$. Using the Dirichlet type 1 integral,

$$E(P_iP_j) = \frac{\Gamma(\alpha_+)}{\Gamma(\alpha_i)\Gamma(\alpha_j)\Gamma(\alpha_{ij})}\cdot\frac{\Gamma(\alpha_i+1)\Gamma(\alpha_j+1)}{\Gamma(\alpha_i+\alpha_j+2)}\cdot\int_0^1(1-\tau)^{\alpha_{ij}-1}\tau^{\alpha_i+\alpha_j+1}d\tau$$

The integrand is the kernel of a beta distribution with parameters $\alpha_i+\alpha_j+2$ and α_{ij}.

$$E(P_iP_j) = \frac{\Gamma(\alpha_+)}{\Gamma(\alpha_i)\Gamma(\alpha_j)\Gamma(\alpha_{ij})}\cdot\frac{\Gamma(\alpha_i+1)\Gamma(\alpha_j+1)}{\Gamma(\alpha_i+\alpha_j+2)}\cdot\frac{\Gamma(\alpha_{ij})\Gamma(\alpha_i+\alpha_j+2)}{\Gamma(\alpha_{ij}+\alpha_i+\alpha_j+2)}$$

$$E(P_iP_j) = \frac{\Gamma(\alpha_i+1)}{\Gamma(\alpha_i)}\cdot\frac{\Gamma(\alpha_j+1)}{\Gamma(\alpha_j)}\cdot\frac{\Gamma(\alpha_+)}{\Gamma(\alpha_++2)} = \frac{\alpha_i\alpha_j}{(\alpha_++1)\alpha_+}$$

To complete the calculation of the correlation, we will also need expectations and variances of P_i and P_j. The marginal distribution of P_i is a beta distribution with parameters α_i and $\alpha_+ - \alpha_i$ (this is derived below), so

$$E(P_i) = \frac{\alpha_i}{\alpha_i + (\alpha_+ - \alpha_i)} = \frac{\alpha_i}{\alpha_+}$$

and

$$\text{var}(P_i) = \frac{\alpha_i(\alpha_+ - \alpha_i)}{(\alpha_i + (\alpha_+ - \alpha_i))^2(\alpha_i + (\alpha_+ - \alpha_i) + 1)} = \frac{\alpha_i(\alpha_+ - \alpha_i)}{\alpha_+^2(\alpha_+ + 1)}.$$

Similarly, $E(P_j) = \frac{\alpha_j}{\alpha_+}$ and $\text{var}(P_j) = \frac{\alpha_j(\alpha_+ - \alpha_j)}{\alpha_+^2(\alpha_+ + 1)}$.

Therefore,

$$\text{cov}(P_i, P_j) = \frac{\alpha_i\alpha_j}{(\alpha_+ + 1)\alpha_+} - \frac{\alpha_i}{\alpha_+}\frac{\alpha_j}{\alpha_+} = \frac{\alpha_i\alpha_j}{\alpha_+}\left(\frac{1}{\alpha_+ + 1} - \frac{1}{\alpha_+}\right) = \frac{-\alpha_i\alpha_j}{(\alpha_+ + 1)\alpha_+^2}$$

and

$$\text{corr}(P_i, P_j) = \frac{\frac{-\alpha_i\alpha_j}{(\alpha_+ + 1)\alpha_+^2}}{\sqrt{\frac{\alpha_i(\alpha_+ - \alpha_i)}{\alpha_+^2(\alpha_+ + 1)}}\sqrt{\frac{\alpha_j(\alpha_+ - \alpha_j)}{\alpha_+^2(\alpha_+ + 1)}}}$$

$$\text{corr}(P_i, P_j) = \frac{-\sqrt{\alpha_i\alpha_j}}{\sqrt{(\alpha_+ - \alpha_i)(\alpha_+ - \alpha_j)}}$$ ■

This expression defines the spurious correlation between the relative frequencies of OTUs *i* and *j*, assuming the original counts are independent gamma random variables with a common scale parameter.

Appendix E. R programs

E.1. Geopolish: Fit the multiplicative model

```
# ------------------------------------------------------------
# Function to iterate multiplicative version of Tukey's
# median polish to fit a multiplicative model to a data matrix.
#  The model is
#  obs = (all effect) * (row effect) * (column effect) * (residual)
#
# Input:
#  df      : data frame containing the named variables
#            'Contig' contains the column identifier
#            'Plot'   contains the row identifier
#            'Count'  is the observed count (must be strictly positive)
#  n.iter  : maximum number of iterations to perform (default 50)
#  toler   : tolerance for stopping criteria (default 10^-9)
#
# Function returns a list with these components:
#  resids  : data frame containing "Contig","Plot","resids"
#  col.eff : a named vector containing column effects for each OTU
#  row.eff : a named vector containing row effects for each Plot
#  all.eff : a scalar for the overall effect
#  msg     : a character string indicating the termination status
# ------------------------------------------------------------
# geometric mean for nonzero entries
  gmean<-function(x) {
  x<-x[x>0]
  prod(x)^(1/length(x)) }
# ------------------------------------------------------------

geopolish <- function (df,n.iter=50,toler=10^-9) {
 df<-df[,c("Plot","Contig","Count")]

 df<-df[df$Count>0,]        # use only positive counts

 end.row.effect<-rep(1,length(unique(df$Plot)))
 names(end.row.effect)<-levels(df$Plot)
 end.col.effect<-rep(1,length(unique(df$Contig)))
 names(end.col.effect)<-levels(df$Contig)
 end.all.effect<-1

 for(kk in 1:n.iter) {
  # sweep rows
  g.row<-tapply(df$Count,df$Plot,gmean)
  resids<-df$Count/g.row[as.character(df$Plot)]
  # sweep columns
  g.col<-tapply(resids,df$Contig,gmean)
  resids<-resids/g.col[as.character(df$Contig)]
  # effects for current iteration
  gg.row<-gmean(g.row)
  gg.col<-gmean(g.col)
  all.effect<-gg.row*gg.col
  row.effect<-g.row/gg.row
  col.effect<-g.col/gg.col
```

```
  # effects for all iterations
  df$Count<-resids
  end.row.effect<-end.row.effect*row.effect
  end.col.effect<-end.col.effect*col.effect
  end.all.effect<-end.all.effect*all.effect

  # stop when all row and col gmeans are == 1
  if (max(abs(g.row-1)) < toler & max(abs(g.col-1)) < toler )
    { names(df)<-c("Plot","Contig","resids")
      msg<-paste("Tolerance met at iteration",kk)
      return(list(resids=df,
                  col.eff=end.col.effect,
                  row.eff=end.row.effect,
                  all.eff=end.all.effect,
                  msg=msg) )
     }
} # end iteration loop

 names(df)<-c("Plot","Contig","resids")
 msg<-paste("Maximum iterations",kk)
 return(list(resids=df,
             col.eff=end.col.effect,
             row.eff=end.row.effect,
             all.eff=end.all.effect,
             msg=msg) )
}
# end function geopolish
```

E.2. Drawdown: Large Count Reduction Algorithm

```
#
# FUNCTION: drawdown (df,sites,contigs,rownums,facts,pct=0.10)
#
# Purpose: To reduce the excessively large individual counts in the dataset
# User supplies
#  df      : data frame containing variables Plot, Contig, Count and
#            additional variables defining the experimental factors
#  sites   : a vector of site values ('Plot') that contain the large counts
#  contigs : a vector of OTU values ('Contig') that contain the large counts
#            (Must be the same length as sites.  Together, sites and contigs
#            uniquely identify the individual large counts that will be
#            examined for potential reduction.)
#  rownums : the data frame row numbers that contain the large counts
#             (used instead of 'sites' and 'contigs' to identify these
#              counts)
#  facts   : a vector of variable names that define the experimental factors
#  pct     : percent separation to maintain between the successive counts
#
# function returns a list with 3 elements
#  * error code
#  * error message
#  * modified data frame, with specified large counts potentially reduced
############################################################

drawdown<-function(df,sites,contigs,facts=NA,targ.rows=NA,pct) {

# initialize
 err.code<-0   # no error
 err.msg<-""   # NULL error message

 # verify the input is valid

 # determine whether user has supplied the actual row numbers
 # or if we need to find the row numbers based on the sites and contigs
 rownums<-targ.rows
 if (length(targ.rows)==1) {
  if (is.na(targ.rows) ){
   # if supplied, vectors 'sites' & 'contigs' should have same length
   if (length(sites) != length(contigs) ) {
    err.code<-1
    err.msg<-"Sites and Contigs must have same length"
    return(list(err.code,err.msg,df))
   }  # endif
  rownums<-c()
  for (kk in 1:length(sites) ){
   rownums[kk]<-which(df$Plot==sites[kk] & df$Contig==contigs[kk],arr.ind=T)
  } # end loop kk
 }  # endif is.na
 }  # endif length

 # variable names in 'facts' should all be in the data frame
 # fact.col = column numbers for the experimental factors
 #            (incl. Plot and Contig)
 fact.col<-which(names(df) %in% c("Plot","Contig"), arr.ind=T)
```

```
 if(!is.na(facts)){
  idx<-which(names(df) %in% facts,arr.ind=T)
  if (length(idx) != length(facts) ) {
   err.code<-1
   err.msg="Not all factor names are in the data frame"
   return(list(err.code,err.msg,df))
  }
 fact.col<-which(names(df) %in% c(facts,"Plot","Contig"),arr.ind=T)
}


 # pct should be between 0 and 1
 if (pct <= 0 | pct>=1) {
  err.code<-1
  err.msg<-"Percent separation must be between 0 and 1"
  return(list(err.code,err.msg,df))
 }

 any.change<-1
 while(any.change==1) {
  any.change<-0 # flag is set to 1 if there are any changes to the counts
  for ( kk in 1:length(rownums) ) {
    rownum<-rownums[kk]
# ; print(paste('rownum:',rownum))
   # get related counts
   values<-df[rownum,fact.col]
         # ; print("values:");print(values)
   tally<-rep(0,length(df$Count))
   for (kk in 1:length(fact.col)) {
    tally.new<-ifelse(unclass(df[,fact.col[kk]])==unclass(values[kk]),1,0)
    tally <- tally+tally.new
    }
   useit<-ifelse(tally>0,TRUE,FALSE)
   counts<-sort(df$Count[useit])
         # ; print("related counts:");print(counts)
    kk <-which.max(counts[counts<=df$Count[rownum]])
    if (kk > 1) {
     if (counts[kk] > ceiling( (1+pct)*counts[kk-1]) )  {
       df$Count[rownum] <- ceiling( (1+pct)*counts[kk-1] )
       any.change<-1
       }
     }
   } #  ; (paste("any.change",any.change))
 }

list(err.code,err.msg,df)
}   # end function 'drawdown'
```

```
#####################################################################
#
# FUNCTION: target.counts(df,pct=0.10,min.count=10,plotit=FALSE)
#
# Purpose: to identify the target counts for potential reduction
# User supplies
#  df        : data frame with variable name 'Count'
#  pct       : percent separation between successive counts
#  min.count : smallest count to be considered for reduction
#  plotit    : logical, should graph be plotted?
#
# Function returns
#  a vector of dataframe row numbers for the counts to be targeted
#
# Counts are flagged if the ratio between the count and next-largest count
# exceeds 1+'pct' and the count is is at least 'min.count'.
# Counts are targeted for potential reduction if their count is
# at least as large as the smallest flagged count.

target.counts<-function(df,pct=0.10,min.count=10,plotit=FALSE,plot.main='') {

 ord<-order(df$Count,decreasing=T)
 ratio<-c(df$Count[ord[-length(ord)]]/df$Count[ord[-1]])
 ratio<-c(ratio,1)
 # min count gets a ratio of 1 -- other ratios are current count to next-
largest count
 target<-ifelse(ratio>(1+pct) & df$Count[ord]>min.count,TRUE,FALSE)
 targ.min<-min(df$Count[ord[target]])
 target<-ifelse(df$Count[ord]>=targ.min,TRUE,FALSE)
 if (plotit==TRUE) {
  plot(df$Count[ord],ratio,
      main=plot.main,
      xlab="Rank of Count",
      ylab="Ratio Successive Counts")
  points(df$Count[ord[target]],ratio[target],bg="red",pch=21)
 }
 ord[target]
}
```