**Introduction**

The species abundance distribution (SAD) is one of the few ecological patterns whose shape is so consistent that it is often considered an ecological law (Lawton 1999, McGill et al 2007). Across varied ecosystems and taxa, the species abundance distribution is nearly always dominated by a few very abundant species and a larger number of increasingly rare species, generating a distinctive hollow- or J-shaped curve (Fisher 1943). Community ecologists have used the species abundance distribution to test numerous theories intended to determine which biological processes are most important for structuring assemblages of species (McGill et al 2007). However, using the species abundance distribution to distinguish amongst theories has proven frustrating because many theories predict similar shapes (McGill et al 2007), and even experimental manipulations generate little variation in the shape of the SAD (Supp and Ernest 2014). After decades of use as a theoretical benchmark, the utility of the SAD for assessing the processes structuring ecological communities is currently unclear.

Increasing evidence for statistical constraints operating on the shape of the SAD (Locey and White, Harte, White et al 2012, older refs – Sugihara?) has complicated our understanding of the SAD as a *biological* pattern. Accumulating evidence suggests that statistical constraints may actually generate the most striking feature of the species abundance distribution – the hollow curve. At its core, the SAD is a representation of how the total number of individuals in a community (N) are partitioned among the number of species (S) in that community. If our implicit null expectation for such a distribution is that is uniformly distributed (i.e. equal numbers of individuals per species), the hollow curve we see in nature is indeed surprising. However, a uniform distribution is not necessarily the appropriate baseline. Just as a Gaussian distribution emerges in the limit of many samples around a population mean, power-law or log-series distributions (i.e. ‘hollow-shaped’ curves) emerge as statistical attractors for abundance distributions (Frank 2009, 2019). Whether using statistical mechanics (i.e. the Maximum Entropy Theory of Ecology (METE); Harte et al 2008, Harte 2011) or combinatorics (i.e. ‘the feasible set’; Locey and White 2013), the random division of individuals into species can, on its own, generate realistic hollow curves. Because these statistically generated SADs are excellent empirical fits to those seen in nature (Harte et al 2011, White et al 2012, Locey and White 2013), it is reasonable to expect that statistical constraints alone may account for the widespread phenomenon of the hollow-curve SAD.

If SADs are statistically inclined to be hollow curves without requiring biological influences, it is no surprise that we have struggled to interpret the hollow curve in biological terms. However, just because the main feature of the SAD may be statistically determined, it does not necessarily follow that the SAD cannot be biologically informative. Biological processes may cause relatively subtle, but meaningful, deviations between observed SADs and their statistical expectations (Locey and White 2013, Harte and Newman 2014). We may be able to use those *deviations* to evaluate theories (Harte and Newman 2014, Xiao et al 2016) byinterpreting simply due to statistical processes operating within the constraints of S and N. Using SADs in this fashion depends on our capacity to detect and quantify deviations between empirical observations and randomness, which requires metrics and computational approaches that allow us to quantify and interpret whatever deviations may exist, across many different types of ecological communities.

In particular, our ability to detect deviations may depend strongly on the size of the community (i.e. its values of S and N). This is because the random expectation for the shape of the SAD, like the random expectations for the large-scale characteristics of many systems made up of numerous subcomponents, emerges from the full array of possible arrangements for the subcomponents, and may be more well-resolved for systems with more subcomponents and therefore more possible arrangements. If the subcomponents of a system fall into an arrangement selected at random from the full array of possible arrangements, and the majority of the possible arrangements share similar large-scale characteristics, the system is likely to fall into an arrangement that also shares these large-scale characteristics. In the case of an SAD with N individuals and S species, the arrangements are the possible divisions of N individuals into S species, and the large-scale characteristic of interest is the shape of the resulting SAD (as captured by metrics such as skewness and evenness). Often, as the number of subcomponents in a system becomes very large, the array of possible arrangements becomes overwhelmingly dominated by arrangements that are very similar to each other in their large-scale characteristics. In statistical physics, information theory, and related fields, where systems often have enormous numbers of subcomponents, this phenomenon is so robust that it forms the basis for the principle of Maximum Entropy and is a powerful tool for inference (Jaynes). In such cases, we can be confident that an observation whose large-scale characteristics deviate even slightly from those shared by the overwhelming majority of possible arrangements is highly unlikely to have occurred at random; the system is subject to a narrow, well-resolved statistical constraint. For the SAD, this would manifest as an extremely narrow and steeply-peaked distribution of values for metrics that describe shape. However, *ecological* systems do not always involve such large numbers of subcomponents. If the number of subcomponents in a system is relatively small, the array of possible arrangements for the system may be more variable in their large-scale characteristics. For the SAD, this would translate into relatively broad distributions of values for the shape metrics. When this occurs, we have less confidence that an observation that differs from the *most* likely characteristics is inconsistent with what would be expected at random. Whether ecological communities are affected by these small-size phenomena, and which community sizes (in terms of S and N) are most affected, is unclear.

Here we set out to establish if, and in what ways, empirical SADs deviate from their distributions of randomly-generated, statistically-constrained SADs. We build upon the combinatorics approach developed by Locey and White (2013). We also examine whether smaller communities indeed have less narrowly-defined statistical baselines than large ones, and whether this appears to affect our capacity to identify deviations. For [# communities] communities spanning X Y and Z taxa, we use combinatorics to characterize the distribution of possible states for the SAD based on the number of species and number of individuals. We then compare *observed* SADs to their corresponding statistical baselines and evaluate 1) if the shapes of observed SADs are consistently unusual given their statistical expectations, 2) how the narrowness of the statistical constraint varies over gradients of *S* and *N*, and 3) whether this variation is associated with variation in whether observations are statistically indistinguishable from random.

**Methods**

*Datasets*

We use a compilation of community abundance data for trees, birds, mammals, and miscellaneous other taxa that has been used in recent macroecological explorations of the SAD (White et al 2012, Baldridge et al 2014). This compilation consists of cleaned and summarized community abundance data for trees obtained from the Forest Inventory and Analysis (ref) and Gentry plots (ref), birds from the North American Breeding Bird Survey (ref), mammals from the Mammal Community Abundance Database (ref), and a variety of less commonly sampled taxa from the Miscellaneous Abundance Database (ref). We filtered these data to remove exceptionally large or small communities, because characterizing the random expectation for the SAD becomes computationally intractable for very large communities, and becomes trivially uninformative for very small ones. We therefore filtered our datasets to remove communities with more than X species or X individuals, or fewer than 2 species or X individuals. We also removed communities for which N = S, because these communities have only one possible SAD. We further filtered the FIA database, which contains roughly 100,000 communities of which approximately 90,000 have fewer than 10 species. Rather than analyze all of these small communities, we randomly selected 10,000 small communities to include in the analysis. All FIA communities with more than 10 species were included in the analysis. Finally, for sites that had repeated sampling over time, we followed White et al (2012) and Baldridge et al (2014) and analyzed only a single, randomly selected, year of data. It should be noted that our analyses include data from the Mammal Community Database and Miscellaneous Abundance Database that were collected over longer timescales and cannot be disaggregated, with an average temporal grain of X. After filtering exceptionally large or small communities and multiple years of data collection, our final dataset consisted of X communities encompassing X taxa, with S and N ranging from 2 X and X to X, respectively (Figure). *Characterizing the statistical baseline*

We use the concept of the *feasible set* to characterize the statistical baseline for the SAD (Locey and White, 2013). For a given number of individuals *N*,there is a finite array of unique ways to partition those individuals into *S* species. The feasible set is the complete set of these unique partitions, or elements. In Locey and White (2013), partitions are unique if and only if they differ in the number of species present with each abundance; neither species nor individuals are distinguishable from each other. Operationally, this means that for *S = 3* and *N = 9*, the vectors *(1, 3, 5)* and *(2, 2, 5)* count as distinct partitions, but *(1, 3, 5)* and *(3, 1, 5)* count as only one element of the feasible set. Alternative assumptions regarding the distinguishability of species and/or individuals would effectively change the probabilities associated with various elements of the feasible set, because some partitions can be achieved via more permutations than others. However, in the absence of strongjustification for one set of assumptions over another, we adopted a simple set of assumptions that have previously been shown to generate realistic statistical baselines (Locey and White 2013).

In the absence of any other process, an SAD with a particular S and N is likely to reflect whatever characteristics are common among the elements of its feasible set. We focus on the shape of the distribution as the characteristic of interest. Metrics related to the shape of the distribution of abundances across species are frequently used in the study of community structure and have been used specifically in the context of distinguishing observed SADs from the feasible set (Locey and White 2013). We focus on two metrics to describe the shape of the SAD, skewness and Simpson’s evenness. Skewness is [….] Simpson’s evenness is a more familiar metric for ecologists. Calculating Simpson’s evenness and skewness for random samples from the feasible set generates distributions of values for each metric that reflect the values that would be likely to occur for an SAD generated at random. Note that skewness [breaks under specific circumstances – s < 3, all abundances equal], and we exclude those cases from analyses of skewness.

Characterizing the feasible set can be computationally intensive, particularly for large combinations of S and N. While it is possible to list all possible partitions of a small number of individuals into a small number of species, the number of elements in the feasible set increases rapidly with S and N and renders it necessary to draw samples from the feasible set, rather than enumerating all of its elements. Unbiased sampling of large feasible sets is itself a nontrivial computational problem, and the computational resources required has constrained previous efforts in this vein (Locey and White 2013). We developed an algorithm to efficiently and uniformly sample feasible sets even for large values of S and N. We implemented this algorithm in an R package, available on GitHub at [www.github.com/diazrenata/feasiblesads](http://www.github.com/diazrenata/feasiblesads); details of the sampling methodology are available at <https://github.com/ha0ye/feasiblesads/blob/algo-vignette/vignettes/sampling_algorithm.Rmd>.

For every community in our database, we uniformly drew 5000 samples from the feasible set to characterize the distribution of statistically probable shapes for the SAD. We filtered the 5000 samples to unique elements. For small values of S and N, it can be impossible or highly improbable to randomly draw 5000 unique samples from the feasible set, but for large communities, all 5000 are usually unique. We then calculated skewness and Simpson’s evenness for every sample from the feasible set, to construct distributions describing the *shapes* likely to occur at random for an SAD with that S and N.

The feasible set’s distributions of Simpson’s evenness and skewness was used to assess whether observed SADs were statistically unlikely given their values of S and N. We calculated Simpson’s evenness and skewness for the observed SADs, and assessed whether each SAD deviated from its feasible set by calculating the percent of values in the corresponding sample distribution less than or equal to the observed value. This percentile rank is comparable across different community sizes, allowing broad-scale assessment across wide ranges of S and N.. If observed SADs simply reflect random draws from their feasible sets, we would expect their percentile rank values to be uniformly distributed from 0 to 100. However, if observed SADs are consistently more skewed or even than their feasible sets, the percentile values will be disproportionately concentrated towards the extremes.

The distributions of summary statistic values from samples from the feasible set also allowed us to describe the relative narrowness or breadth of the statistical constraint over gradients of S and N. We quantified the narrowness of a distribution as the ratio of the range of values encompassed within a one-sided 95% interval to the full range of values in the distribution ([Figure](#_Figure_2:_95%)). This allowed us to isolate and compare the *narrowness* of distributions that are not necessarily normally distributed and vary considerably in their actual values, means, ranges, and standard deviations. This metric corresponds qualitatively to more computationally-intensive approaches to measuring the self-similarity of the elements of feasible sets (Supplement).

**Results and Discussion**

We found evidence that the shapes of the SADs for a range of real ecological communities are statistically unlikely when compared to their feasible sets. For four of the five datasets we analyzed – BBS, Gentry, Mammal Communities, and Misc. Abund – empirical SADs are highly skewed and highly uneven relative to their feasible sets much more frequently than would be expected by chance ([Figure](#_Figure_3:_Skewness_1), [Figure](#_Figure_4:_Evenness)). Combined across these four datasets, 16% of observed SADs are more skewed than 95% of their feasible sets, and 31% are less even than 95% of their feasible sets. By chance we would expect only 5% of observed distributions to fall in these extremes. However, we ), for which Why the FIA communities show a different pattern is unclear, but may be partially due to their small size, which we explore further below. For the four other datasets, however, our results suggest that nonrandom processes drive observed SADs to be less even than would occur by chance. This may be the signature of ecological mechanisms operating on top of the statistical constraint.

Some of the variation across our datasets in how communities compare to their feasible sets may be due to statistical issues related to community size. Community size – in terms of *S* and *N* – may affect our ability to distinguish between deviations and randomness via its potential effect on how similar the elements of the feasible set are to each other. The FIA communities are the smallest in our database, and we saw relatively weak evidence of these communities deviating from their feasible sets. Similarly-sized communities from the other datasets show similarly weak, or absent, evidence for deviations. This would be consistent with the possibility that small communities may not have enough possible arrangements if their subcomponents, or elements in their feasible sets, to generate highly resolved distributions for the most probable shapes.

When we compared the distributions of shape metrics for small communities to those for large ones, we samples from the feasible sets for small communitiesgenerate

**Conclusion**

In general, our results suggest

that the shape of the SAD is not entirely a statistical artefact – that there are indeed biological processes that generate a particular, highly uneven, form for many empirical SADs. Identifying the processes that generate these deviations in ecological communities will require further exploration. One logical next step is to test whether existing theories (e.g. neutral theory, METE, ???) and common functional approximations (logseries, exponential, and log normal) can accurately predict deviations from the feasible set. Our results also suggest that the prevailing processes that structure communities tend to be ones that push abundance distributions towards a more uneven state, rather than those that cause individuals to be spread more evenly across species.

While our results provide support for ecological processes structuring species abundance distributions, they also suggest there may be limits to our ability to distinguish between communities that could be structured effectively at random and ones that are likely structured via ecological process – particularly for small communities. Within our datasets, the FIA communities seemed to have especially broad distributions of shapes represented in their feasible sets. These communities range in size from x to y species and x to y individuals. While these are by no means hard thresholds, they may indicate a general range of values below which we have relatively diminished power to distinguish deviation from randomness. Unless we can develop more sensitive methods for identifying deviations even for these small communities, we may stand to learn the most by focusing on SADs from relatively large communities. In the meantime, sampling the range of forms represented in the feasible set helps us identify when the distribution of shapes present is relatively broad, and may even help us develop a type of power analysis to find the smallest-detectable deviation for a given community size.

It is also important to recognize that there are multiple plausible approaches to defining the statistical baseline for the SAD, of which we have taken only one. Other formulations for the statistical baseline may be equally valid and can generate different statistical expectations, including forms that approximate exponential, Poisson, or log-series distributions. There is currently no unambiguous logical argument for one baseline over another, but comparing the performance of different baselines is clearly an important next step in this process towards reinvigorating the use of the SAD as a diagnostic tool.

Characterizing and adjusting for a statistical baseline for the SAD, as we demonstrate here, refreshes our perspective on the distribution and opens up several new avenues for better understanding how and when biological drivers affect its shape. Persistent deviations between observed communities and their baselines may be evidence of biological processes operating on top of fundamental statistical constraints, and focusing on these deviations could offer new leverage for evaluating theoretical predictions for the SAD. In doing so, we must appreciate that there is considerable nuance to defining the appropriate statistical baseline and calibrating our expected power to *detect* deviations, especially for small communities. Exploring other constructions for the baseline, and developing methods for establishing if not improving the constraints on our ability to detect deviations, would further clarify how statistical constraints manifest in the SAD and what power we have to disentangle biological signal from randomness. Our results here suggest that statistical constraints have strong effects on the SAD, but that these constraints alone do not fully account for the extremely uneven SADs we observe in nature – leaving an important role for ecological process. Continuing to explore and account for the interplay between statistical constraint and biological process constitutes a promising and profound new approach to our understanding of this familiar, yet surprisingly mysterious, ecological pattern.

**Figures**

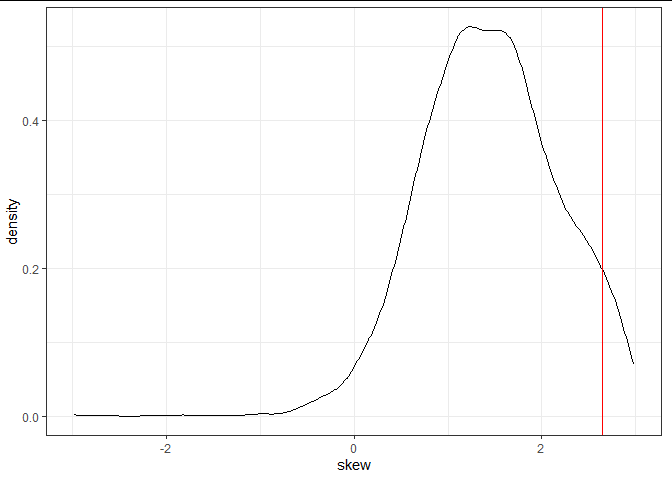
##### Figure 1: Communities by dataset, S, N

##### 

##### Figure 2: 95% ratio schematic

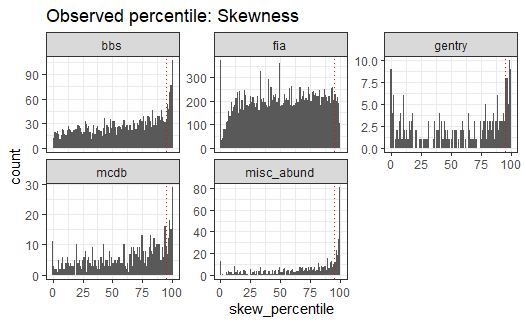
Range of values for 95% density

÷



Full range of values

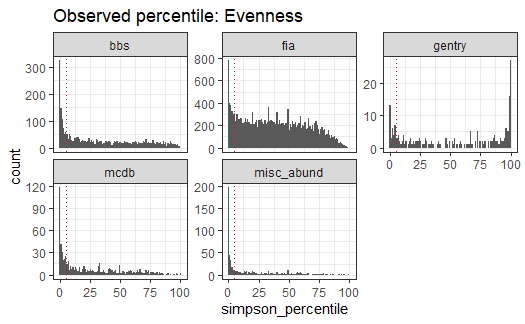
##### Figure 3: Skewness percentile results



Dotted line is 95.

Excludes communities for which S < 3

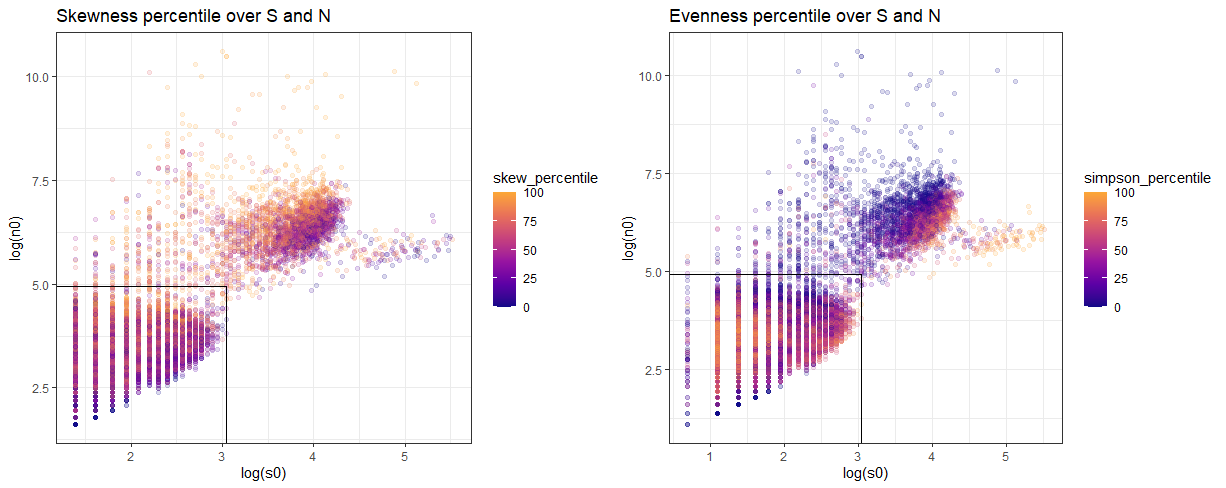
##### Figure 4: Evenness percentile results



Dotted line is 95.

##### 

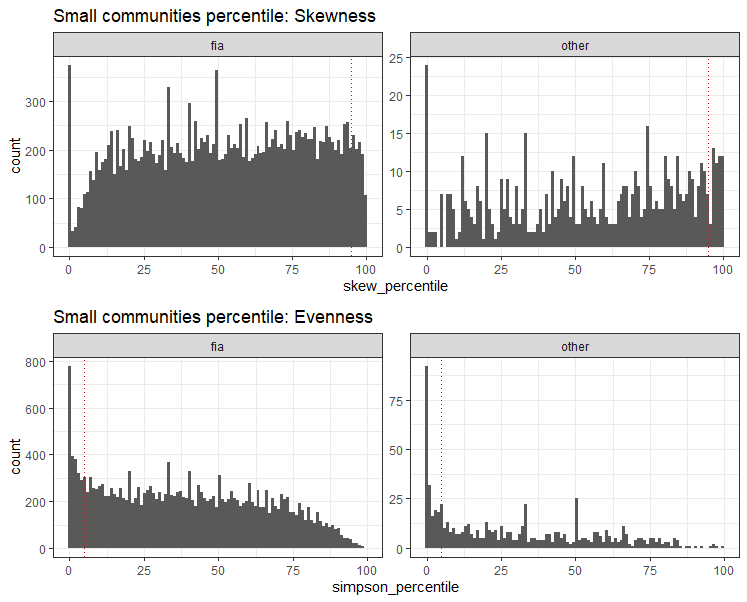
##### Figure 6: Percentile scores over S and N



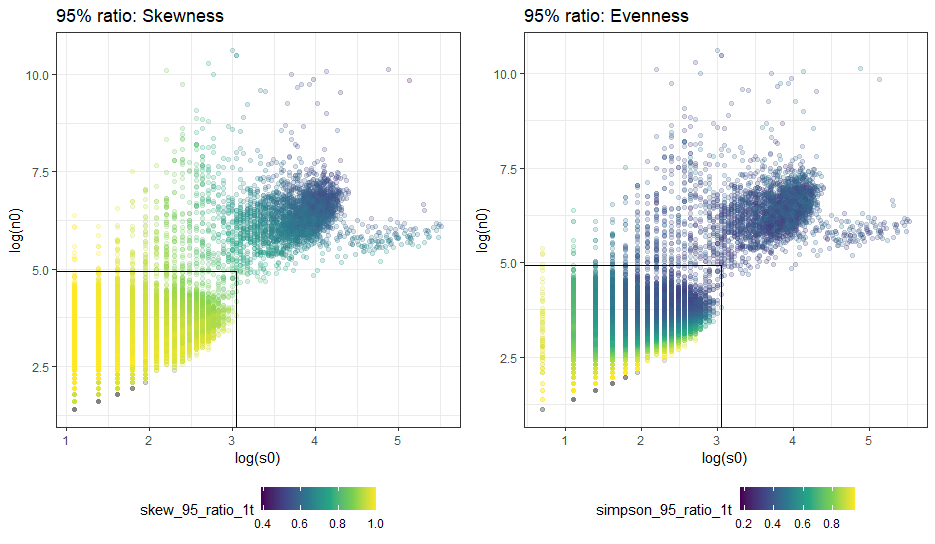
Skewness excludes s < 3

Black line outlines FIA maximum s0 and n0

##### Figure 7: Percentile outcomes for small communities



##### Figure 8: 95% ratio over S and N



Skewness excludes s < 3

Black line is FIA maxima

**Supplement stuff**

*Effects of rarefaction*

Generally, it is logistically impossible to conduct exhaustive surveys of ecological communities, and even the most intensive surveys of nature are likely to miss some species – particularly rare or cryptic ones. To test whether missing rare species influenced the relationship between observed SADs and their feasible sets, we ran our analyses on both raw data and on SADs adjusted via rarefaction. For each raw SAD, we took the mean number of estimated species from [] richness estimators, and added species to reach this mean estimated richness. We reasoned that species missed during sampling are likely to be rare, and added a single individual each for each added species. This also allowed us to explore the consequences of rarefaction while making the smallest possible changes to S and N.