**Introduction**

The species abundance distribution (SAD) is one of the very few ecological patterns whose consistency in shape makes it akin to 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). Tempted by this remarkably general pattern, community ecologists have used the species abundance distribution to test numerous theories designed to determine which biological processes are most important for structuring assemblages of species (McGill et al 2007). However, using the species abundance distribution to infer underlying processes has proven frustrating because there is so little variation in the shape of this distribution, even in response to experimental manipulations (Supp and Ernest 2014), and many theories produce similar predictions that perform comparably well when compared to observed distributions (McGill et al 2007).

Ecologists have generally focused on explaining the species abundance distribution as emerging from ecological processes, but *statistical* processes may have surprisingly powerful contributions as well. Indeed, accumulating evidence suggests that statistical constraint may play a large role in generating 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 and begs for explanation. However, analogous to the way 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). In ecology, different approaches using statistical mechanics (i.e. Maximum Entropy Theory of Ecology (METE); Harte et al 2008, Harte 2011) or combinatorics (i.e. ‘the feasible set’; Locey and White 2013) have shown that just using a community’s values of S and N, statistical processes alone can generate hollow curves that are excellent empirical fits to those seen in nature (Harte et al 2011, White et al 2012). Therefore, we have good reason to expect that based on statistical constraints alone, most SADs will be hollow curves.

If SADs are statistically inclined to be hollow curves before biology even enters the picture, it is no surprise that we have struggled to interpret the hollow curve in biological terms. It may be more informative to evaluate SADs not in terms of their absolute shape, but in terms of their shape relative to what we would expect simply due to S, N, and the statistical constraint(Frank 2009, Locey and White 2013). Biological processes may cause relatively subtle, but meaningful, deviations between observed SADs and their statistical expectations (Locey and White 2013, Harte and Newman 2014) and we may be able to use those deviations to evaluate theories (Harte and Newman 2014, Xiao et al 2016). However, this depends on our capacity to detect deviations between empirical observations and randomness. This capacity depends strongly on the distribution of statistically-likely outcomes, which in turn depends on the size of the system in terms of S and N. Generally, in the study of complex systems, predictions for the most-likely state of a system are most easily and confidently derived “in the limit” of aggregating over large numbers of components – for example, atoms of gas in a vessel (Frank 2009, Harte et al 2011). As the number of particles in the system becomes very large, the set of *possible* states of the system clusters tightly around a single highly-likely state. (Jaynes 1957). Unlike vessels of gas, ecological communities often do not have effectively infinite subcomponents. Such small systems may have a relatively broad distribution of likely possible forms, in which case even observations that are quite different from the *most* likely outcome are not necessarily highly unlikely. When this occurs, we cannot be sure if an observation that differs from the most-likely outcome, but falls within the realm of statistical possibility, is a result of random sampling or a reflection of nonrandom process. Although we expect this phenomenon to be most relevant to small systems, *a priori* we do not know if ecological communities suffer from this issue, nor what the specific ranges of S and N may be that create this issue.

Here we set out to establish if, and in what ways, we can detect consistent deviations between empirical SADs and their statistical constraints. We also examine whether small communities are more likely to exhibit a lack of detectable deviations between an empirical SAD and its distribution of possible shapes, which could limit our ability to make confident statistical distinctions in those contexts. For [# communities] communities spanning X Y and Z taxa, we characterize the distribution of statistically-likely forms 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) whether 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 seems to impact whether we can detect deviations.

**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 using the feasible set approach (White et al 2012, Baldridge et al 2014). It includes data from trees in the Forest Inventory and Analysis and Gentry plots, the North American Breeding Bird Survey, the Mammal Community Abundance Database, and a variety of less commonly sampled taxa in the Miscellaneous Abundance Database.

For datasets with observations from multiple years, we follow White et al (2012) and Baldridge et al (2014) and analyze only a single year of data. The Mammal Community Abundance Database and Miscellaneous Abundance Database include data collected over longer timescales that cannot be disaggregated, with an average temporal scale of X.

Our approach to characterizing the statistical constraint on the SAD becomes computationally intractable for the very largest communities, and becomes trivially uninformative for the very smallest communities. 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.

The FIA database contains roughly 100,000 communities, of which 90,000 have fewer than 10 species and X individuals. Rather than analyze all of these small communities, we randomly selected 10,000 small communities to include in the analysis. We analyzed all communities in FIA with more than 10 species. In all, we analyzed X communities encompassing X taxa, with S and N ranging from 2 to X and X to X, respectively.

*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 possible set of unique ways to partition those individuals into *S* species. The feasible set is the complete set of these unique partitions. The distribution of shapes or statistical properties of the elements of the feasible set constitutes a statistical baseline for the SAD. In the absence of any biological process, a randomly generated SAD is likely to reflect the statistical characteristics most common in the feasible set. To assess whether observed SADs deviate from this baseline, we compared the shapes of empirically-observed SADs to the distributions of shapes present in their feasible sets. We used the range of variation in shapes represented in the feasible set to explore how the narrowness of the constraint on the probable shape of the SAD - which may modulate our power to detect deviations - varies over gradients of S and N.

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>. Importantly, for the purposes of defining unique elements of the feasible set, neither species nor individuals are considered distinguishable from each other (Locey and White, 2013). Partitions are unique if and only if they differ in the number of species present with each abundance. That is, 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. Each distinct element of the feasible set is considered equally probable to occur at random. Alternative assumptions regarding the distinguishability of species and/or individuals would 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 approximate reality (Locey and White 2013).

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.

For each set of samples from a feasible set, we calculated the skewness and Simpson’s evenness for every sample. The distribution of values for these summary statistics reflects the likely statistical characteristics for a random SAD with that S and N, and the range of variation in this distribution shows how narrowly a feasible set clusters around a particular shape. Note that skewness [breaks under specific circumstances – s < 3, all abundances equal], and we exclude those cases from analyses of skewness.

To compare observed SADs to their feasible sets, we compared the values of the summary statistics calculated for the *observed* SAD to the distribution of summary statistics obtained from samples from the corresponding feasible set. We used the percentile rank, calculated as the percent of values in the sample distribution less than or equal to the observed value. At random, percentile rank values are uniformly distributed from 0 to 100. 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 the one-sided 95% interval to the full range of values in the distribution.

*Effects of rarefaction*

Whenever we survey the species present in a real system, we are likely to miss some species – particularly rare or cryptic ones. To test whether this affected our results, 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.

**Results and Discussion**

We detected appreciable differences between empirical SADs and their feasible sets, with a range of variation that may be a promising source of new information and statistical leverage. We observed these deviations more consistently for large communities than for smaller ones – in particular, the very small communities that comprise most of the FIA dataset - and we suggest that the range of variability of the feasible sets of small communities limits our ability to detect deviations. Adjusting for rarefaction increased the signal of deviation.

*Empirical SADs compared to their baselines*

For four of the five datasets we analyzed – BBS, Gentry, MCDB, 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. 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. Adjusting for rarefaction increased the strength of this signal: for these datasets, 18% of adjusted SADs are more skewed than 95% of their feasible sets, and 38% are less even. In nearly every case, the percentile scores for rarefaction-adjusted SADs were more extreme than for the raw SADs. If anything, incomplete sampling seems to cause us to *under* estimate the deviation between observation and randomness.

This may be evidence 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. So far, we have struggled to understand or even identify these processes because we have focused on the wrong aspects of this form. We can use the deviations between empirical SADs and their feasible sets as a new source of leverage for fitting models and evaluating theories. A model that deviates from the feasible set consistent with observed distributions will be much more convincing than one that simply predicts the central tendency of the feasible set, even though both models will predict plausible-seeming hollow curves. A logical starting point will be to test predictions from established theories (e.g. neutral theory, METE, ???) and common functional approximations (logseries, exponential, and log normal) for the SAD to evaluate which ones make accurate predictions regarding deviations.

While there is an overall signal of high unevenness in these four datasets, there is also considerable heterogeneity in how empirical SADs compare to their feasible sets. Some of this variation may be statistical, driven by special circumstances regarding S and N. For example, certain Gentry communities have very low average abundances, which forces all elements of the feasible set to be fairly even; coincidentally, these are the only group of communities for which empirical SADs are unusually *even* relative to their feasible sets. However, we also see considerable variation between communities with similar values of S and N. More focused comparisons between communities may show whether there are identifiable differences between these communities that systematically cause some to deviate and some not – differences to which we would not otherwise have been attuned.

We detected considerably less pronounced deviations for communities from the FIA communities. For FIA, percentile scores were near uniformly-distributed for skewness (5% of observations are more skewed than 95% of the feasible set), and much noisier than any of the other datasets for evenness (11.5% of observations are less even). This may be primarily an effect of small community size: the FIA communities are the smallest in our database. There was not an obvious difference between FIA communities and comparably-sized communities from other datasets in the distribution of percentile values – although note that the vast majority of communities in this size range are FIA.

*Small community effects*

Community size – in terms of *S* and *N* – may affect our ability to distinguish between deviations and randomness via its effect on the variability of forms represented in the feasible set. We found that the feasible sets for small communities have broader distributions of evenness, and especially skewness, than those for large communities. For communities of the sizes represented in the FIA dataset, the 95% interval of skewness values often encompasses more than 80% of the entire range of values; for larger communities, the 95% interval spans closer to 60% of the full range. This is consistent with concepts from statistical mechanics. Large communities have many components that can be arranged in many ways, and most of these arrangements cluster around a relatively specific highly-likely state. If an observation differs even a small amount from this most-likely state, it is readily detectable as highly unlikely to have occurred by chance. In contrast, small communities have relatively few possible arrangements and relatively broad distributions of likely shapes. Observations may deviate from the *most*-likely form, but only the most extreme deviations will be highly unlikely given the breadth of the corresponding probability distribution.

These small-community considerations appear to be relevant for ranges of S and N that are quite common in ecology. The FIA communities range in size from x to y species and x to y individuals. These are by no means hard thresholds, but they do 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 in these small communities, we 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 provides context to identify when the breadth of variation in the feasible set is so great it affects our power to detect deviations.

*Alternative statistical baselines*

There are multiple plausible approaches to defining the statistical baseline for the SAD, of which we have taken only one. We have taken the feasible set defined as the set of possible unique, unordered divisions of *N* interchangeable individuals into *S* interchangeable species. This is a simple definition that has been previously shown to generate a statistical baseline that falls close to, but not exactly matching, observed SADs (Locey and White, 2013). However, 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. In the absence of an unambiguous logical argument for one baseline over another, we suggest performing analyses similar to the one presented here using alternative baselines.

**Figures**

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





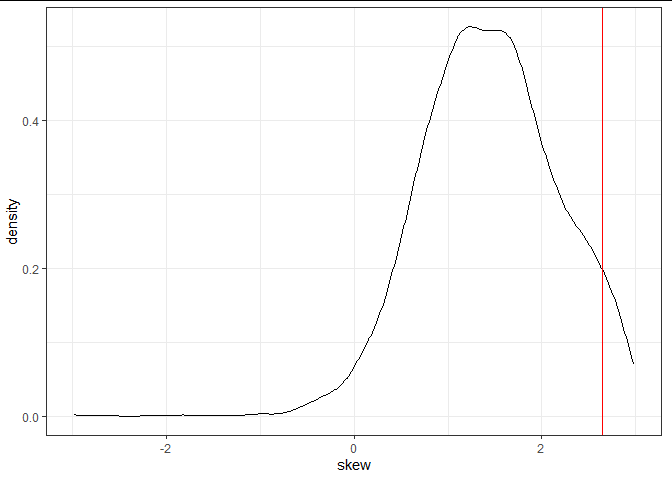


##### 

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

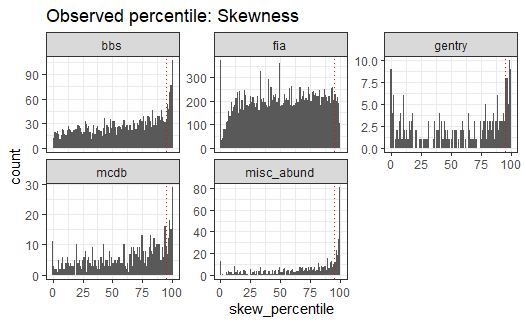
Range of values for 95% density

÷



Full range of values

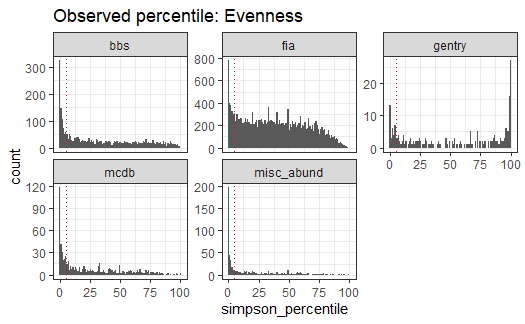
##### Figure 3: Skewness percentile results



Dotted line is 95.

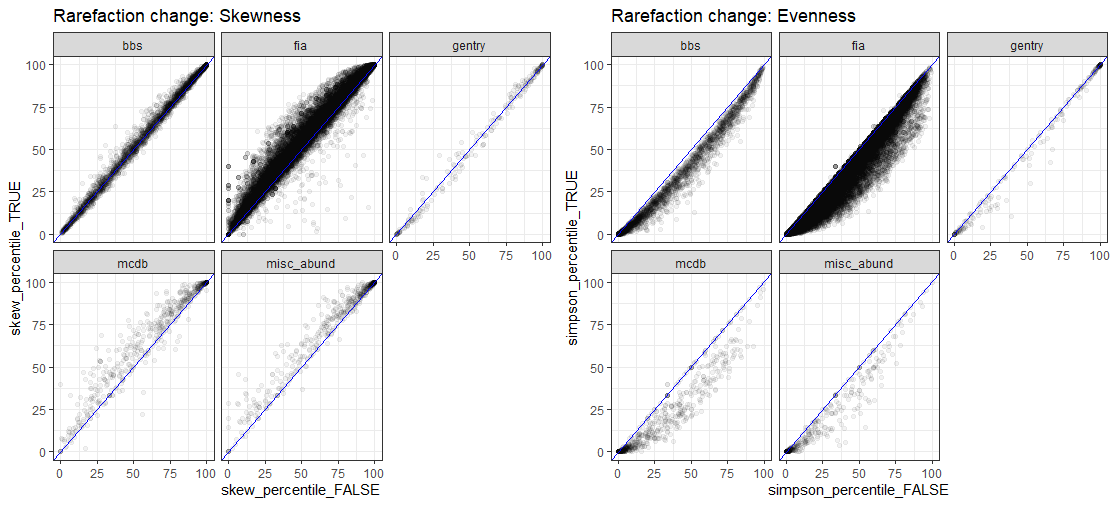
Excludes communities for which S < 3

##### Figure 4: Evenness percentile results



Dotted line is 95.

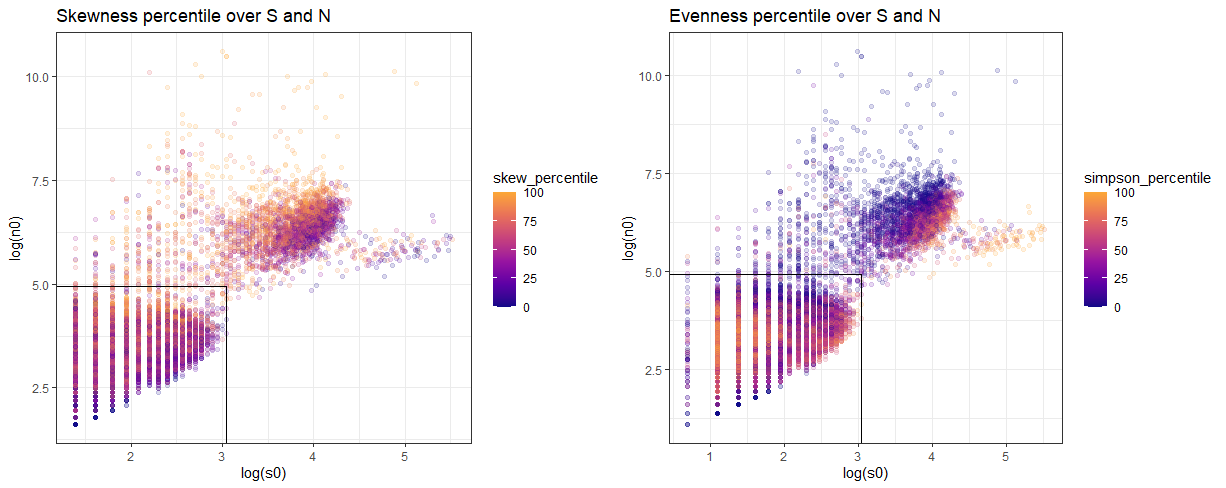
##### Figure 5: Rarefaction effects



Blue line is 1:1

Skewness excludes s < 3

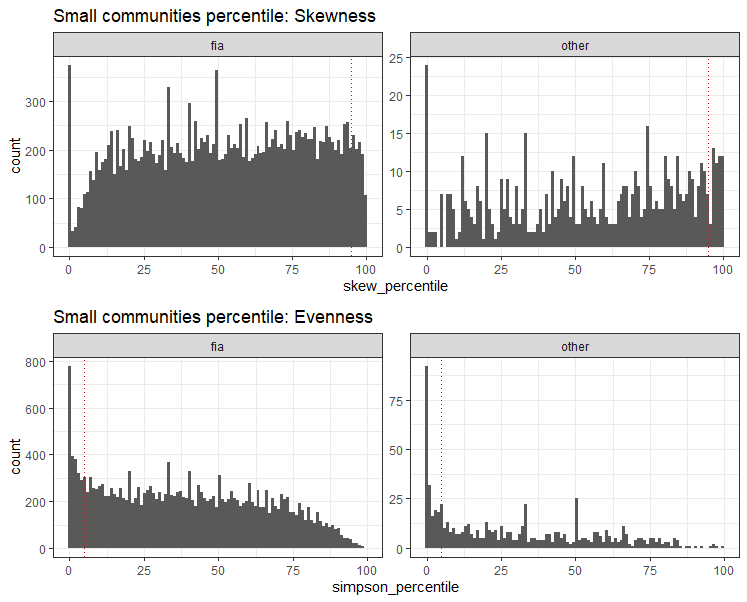
##### Figure 6: Percentile scores over S and N



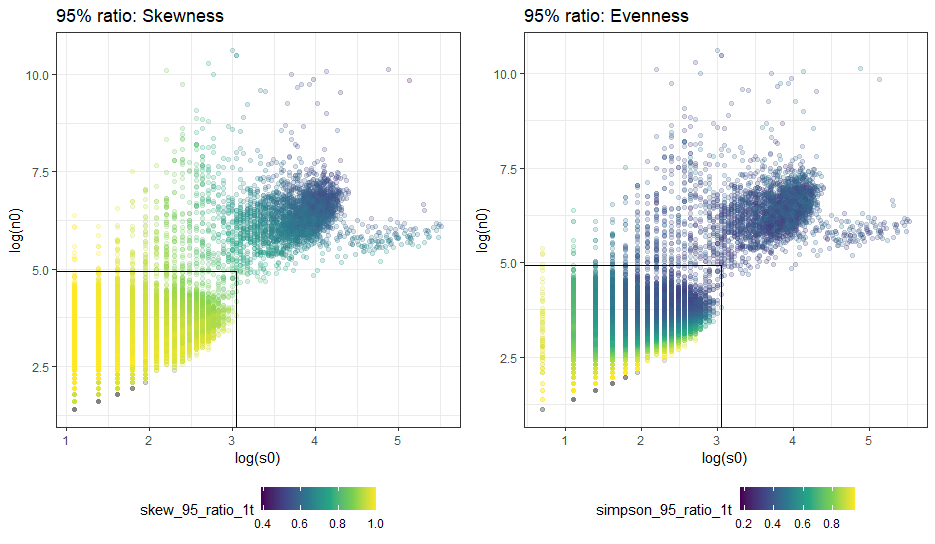
Skewness excludes s < 3

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