**Introduction**

The species abundance distribution (SAD) has been a major focus of both theoretical and empirical traditions in community ecology, largely because it is one of the very few ecological phenomena to conform to a consistent pattern approaching a law (Lawton, McGill, …). Regardless of the biological specifics of an assemblage, it will almost certainly be dominated by a few very abundant species, with relatively many rare species rounding out the abundance distribution (). Tempted by this remarkably general pattern, community ecologists have developed an array of theories to explain the underlying biological processes, as well as methods to divine the internal state of a system based on subtle variations in the SAD (). Ironically, this same generality of form has confounded our efforts to extract biological information from the SAD. There is strikingly little variation in form among empirical SADs, even in systems subject to experimental manipulations (Supp, ...). Most theories produce similar predictions and perform comparably well when compared to observed distributions, making it difficult to use the SAD to evaluate competing theories (McGill).

Multiple lines of recent reasoning suggest an explanation and perhaps a solution for this puzzle. The SAD seems subject to a statistical constraint that encourages it to be a hollow curve, even in the absence of any biological mechanism. The hollow curve is intuitively striking, but focusing on what may be a statistical inevitability warps our perception of the important aspects of the pattern. Frank () used a statistical mechanics framework to show that hollow curves, including a log-series – one of the most common mathematical approximations for empirical SADs – can easily be obtained as statistical by-products of [RMD forgets the precise process and needs to revisit Frank’s stuff]. Harte and associates () used the principle of entropy maximization to find the most-likely form for numerous ecological distributions, given sparse information about broad community properties and minimal assumptions regarding ecological mechanism. They also found that a log-series emerges as the most-likely form for the SAD (). This log-series is an excellent fit to empirical SADs (White et al 2012). Finally, Locey and White (2012) put forward an entirely non-biological derivation for the statistical constraint on the SAD. They leveraged the fact that there are a finite number of ways to divide *N* individuals among *S* species, and used combinatorics to characterize the range of *possible* forms for SADs, or feasible set, for realistic values of S and N. They showed that practically all feasible sets have strong central tendencies towards hollow curves ().

If, as seems to be the case, SADs are statistically driven 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. But there is hope for a way forward: if we recognize and account for the statistical constraint, we may be able to isolate subtler but more biologically meaningful variation in SADs. Specifically, we can evaluate SADs not in terms of their absolute shape, but in terms of their shape *relative to what we would expect simply due to statistical constraint* (). If empirical SADs consistently deviate from their statistically most-likely forms, these deviations may signal biological processes operating over and above the statistical constraint (). We can then evaluate theories in terms of how well they predict the *deviations,* or how much additional predictive power they give us over and above what is inherent to S and N(). If we cannot distinguish between observations their statistically-most-likely forms, there may be little information we can extract from the SAD alone – because no processes consistently regulate the SAD, because so many processes operate in various directions that the emergent distribution is effectively random, or because the statistical peculiarities of the system make it difficult to disentangle signal from noise ().

Locey and White () demonstrated a blueprint for evaluating SADs in the context of their statistical constraints. They drew samples the *feasible set* and compared observed SADs to the central tendencies of their feasible sets. Locey and White’s initial efforts showed that empirical SADs are often more skewed and less even than the central tendencies of their feasible sets.

This is an especially appealing approach because the feasible set is a truism based on S, N, and combinatorics; we need not make introduce additional assumptions or mathematical machinery. It also highlights an inherent challenge for interpreting results in terms of their statistical baselines. The number of elements in the feasible set, and their general characteristics, such as evenness or skewness, depend strongly on S and N. In general, and especially for communities with high S, N, or average abundance, the feasible set is large and has a well-defined central tendency we can interpret as the statistical constraint. However, some combinations of S and N will produce feasible sets that are very small and/or have a relatively weak central tendency. If the range of possible forms for the SAD is broad and relatively evenly distributed, it may not be possible to detect meaningful deviations. This confounds our efforts to distinguish between observations and the central tendency, but also to evaluate *theoretical* predictions for the SAD in terms of how well they can pinpoint the real SAD over and above what is forced by S and N. We can confidently assume that *some* ecological communities will fall into this category, although Locey and White may not have dealt with them directly because they restricted their analysis to communities with more than 10 species. It is therefore important to both quantify whatever deviations we can detect from the statistical constraint, and establish expectations for when the central tendency will be too vague for us to disentangle signal from noise.

Here, we extend Locey and White’s approach to 1) more specifically establish the ranges of S and N for which we can confidently detect deviations from a strong central tendency and 2) quantify the extent and ways in which empirical SADs deviate from their feasible sets.

**Methods**

*Datasets*

We use a compilation of datasets of community abundances for trees, birds, mammals, and miscellaneous other taxa that has been used in recent macroecological explorations of the feasible set (White et al 2012, Baldridge et al 2014). It includes data from trees in the Forest Inventory and Analysis and Gentry plots, birds in the North American Breeding Bird Survey, the Mammal Community Abundance Database, and diverse less commonly sampled taxa in the Miscellaneous Abundance Database.

For datasets with observations from multiple years, we followed 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.

Sampling the feasible set is impractical for very large or very small communities. It is computationally intractable to sample the feasible set for very large communities, and very small communities have too few elements in their feasible set for our approach to be truly applicable (see below). The size of the feasible set varies as a function of S, N, and the *ratio* of N to S; communities with many individuals, many species, and a high ratio of N:S have the largest and least tractable feasible sets. We therefore remove communities with more than X species or X individuals, or fewer than 2 species or X individuals. Additionally, the FIA database contains roughly 100,000 communities. Of these, approximately 90,000 have fewer than 10 species and X individuals. Rather than analyze all of these small communities (which, again, requires nontrivial computational resources), we randomly selected 10,000 small communities to include in the analysis. In all, we analyzed X communities encompassing X taxa, with S and N ranging from 2 to X and X to X, respectively.

*Characterizing the feasible set*

Enumerating and uniformly sampling the elements in the feasible set is a nontrivial challenge in number theory and combinatorics, and past efforts to do so have been highly computationally intensive. We developed an algorithm to efficiently and uniformly sample feasible sets for even comparatively large S and N. We implement this algorithm in an R package, available on GitHub at [www.github.com/diazrenata/feasiblesads](http://www.github.com/diazrenata/feasiblesads).

[explanation of sampler]

*Effects of S and N on the size and central tendency of the feasible set*

S, N, and the ratio of N to S determines both the number of elements in the feasible set and the properties of those elements. This may affect our ability to detect or draw meaning from deviations between an SAD of interest and its feasible set.

In a purely statistical null scenario, every possible unique element of a feasible set is equally likely to be the one we happen to observe. As Locey and White (2012) demonstrated, many feasible sets have a strong central tendency: most of their unique elements share a similar form. Happening upon a SAD that does not match this general form would be highly surprising under the null scenario, and might betray a process driving the system away from its most-likely state.

This power of this approach to detect deviations depends on the narrowness of the central tendency of the feasible set. The most informative findings will be ones where we find observed SADs fall outside of densely concentrated central tendencies: in these cases, despite an overwhelmingly specific statistical constraint, the observed distribution adopts another form. If a feasible set does not have a strong central tendency, it will be difficult to distinguish between deviations and randomness. It is safe to assume that at least some combinations of S and N will have such vague central tendencies that we cannot detect deviations that might signal a nonrandom process.

While we do not know, *a priori*, which ranges of S and N will have sufficiently narrow central tendencies, we can establish how the strength of the central tendency varies across combinations of S and N and check whether the specificity of the central tendency seems to inform our ability to detect deviations.

We measured the self-similarity of elements of the feasible sets for combinations of S and N spanning the ranges represented in our compilation of datasets. For each S and N combination, we drew a maximum of 4000 unique samples from the feasible set. We then randomly drew a maximum of 4000 pairs of SADs from each set of samples and calculated four measures of similarity between each pair: a direct calculation of R2 common in macroecology, the coefficient of determination from a linear model fitting one sample to the other, the proportion of individuals allocated to different species, and the K-L divergence. Each of these measures is imperfect for these purposes, but the four measures all give qualitatively the same outcomes. Note that, for small communities, both the number of unique elements in the feasible set and the number of possible pairwise combinations of elements can be considerably lower than 4000. We take the mean value of each of these measures as a rough measure of the specificity of the statistical constraint for that community.

*Comparing observed SADs to their feasible sets*

We compare the forms of observed SADs to their feasible sets by establishing a distribution of summary statistics for the SADs present in the feasible set and comparing the summary statistic values for the observed SADs to these distributions. We used skewness (e0171::skewness) and Simpson’s evenness (vegan::diversity(index = “simpson”)) as our summary statistics, and calculated the percentile rank of the value for each observed SAD compared to the distribution of values obtained from its feasible set. Skewness can be calculated multiple ways, but using different formulas yielded qualitatively identical results. It cannot be calculated for SADs with fewer than three species, and it is infinite in the special case where every species has the same number of individuals. Simpson’s evenness does not have these idiosyncrasies.

The range of values for both statistics depends on S and N, so in order to synthesize across communities with different S and N we must focus on the percentile value, rather than on the raw values of the statistics. We calculated percentile rank as the percent of values in the distribution less than or equal to the observed value. Note that the precise meaning of this value differs subtly depending on whether the distribution of samples from the feasible set represents a true distribution of *samples* or an exhaustive account of all elements in the feasible set. If the samples are exhaustive, the percent of samples <= to the observed value may be 100, but can never be 0 – even if the observed value is the lowest value possible, it is still <= itself. However, if the samples are not exhaustive, we may obtain a percentile value of 0 if the observed value is less than the lowest value we came across via sampling.

At random, we would expect the percentile values to be uniformly distributed from 0 to 100. If observed SADs are consistently more skewed or even than their feasible sets, the percentile values will be nonuniform. If, as we suspect given our reasoning about the central tendency (above), smaller communities have less defined central tendencies, we may not detect consistent deviations from randomness in those communities. We therefore present results for the distributions of percentile values both for all communities, and broken out according to the size of the feasible set.

*Effects of rarefaction*

Whenever we sample the species present in a real system, we are likely miss some species – particular rare or cryptic ones. To test whether adding these presumably present-but-unaccounted-for species affects our results, we ran our analyses on both the abundance data as reported and on SADs adjusted via rarefaction. For each raw SAD, we took the mean number of estimated species from X Y and Z richness estimators, and added species to reach this mean estimated richness. We assumed that the missing species are likely to be rare, and added a single individual of each added species. This allows us to explore the consequences of rarefaction while making the smallest possible changes to S and N.

**Results**

*Effects of S and N on the feasible set*

The number of elements in the feasible set increases rapidly with both S and N. Most feasible sets exhibit a discernable central tendency, which becomes most pronounced for very large feasible sets.

*Observed SADs relative to their feasible sets*

Overall, we see a disproportionate number of high scores for skewness and low scores for evenness for observed SADs compared to their feasible sets. However, this result is powerfully contingent on the size of the feasible set. Communities whose feasible sets have fewer than ~e^10 elements do not seem to have percentile scores that are discernably non-uniform in their distribution, but communities with larger feasible sets are decisively non-uniform.

**Scrap paragraphs**

Borrowing insight from the principle of maximum entropy (), we can expect the strength of the central tendency, and our capacity to detect meaningful deviations, to depend This insight echoes the principle of maximum entropy, well-established in statistical mechanics (Jaynes, Harte). For a system with numerous subcomponents that are arranged into various configurations – for example, atoms of a gas held in a box - the most likely state of the system in aggregate is the one that can be achieved via the most possible arrangements of the subcomponents. As the number of subcomponents in the system approaches infinity, the vast majority of possible arrangements grow so similar that is overwhelmingly unlikely for the system to fall into any other state.