**Introduction**

Over decades of ecologists’ theorizing, the species abundance distribution has proven a tempting but surprisingly slippery source of information about the processes structuring ecological communities. Almost all communities have a large proportion of rare species, leading to classic “hollow curve” SADs. Generations of ecologists have attempted to use this remarkably general pattern to evaluate theory and make inference about communities’ state and structuring forces. These efforts have been frustrated, at least in part, by the same generality that makes the SAD so intriguing: most theories also predict hollow curves, and the range of variation in the form of empirical SADs is small.

Recent work by Locey and White (2013) offers an explanation for, and potential way out of, this predicament. They demonstrated that hollow curves dominate the range of *possible* forms, or “feasible set”, for the SAD. Simply due to the mathematical constraints imposed by partitioning a particular number of individuals into a particular number of species, the most-likely form for the SAD is a hollow curve. The hollow curve itself is not necessarily a symptom of any additional process, and efforts to reproduce or explain the hollow curve may not yield compelling insights into *ecology*.

However, we may be able to use the feasible set, or range of possible forms, for the SAD as a baseline against which to evaluate real SADs. If an empirical SAD is unusually even or skewed compared to its feasible set, this might indicate that biological forces have driven that community away from its most-likely, random, state. If such deviations are consistent across communities, we may be able to use the *deviations* as the biological signature around which to develop and test theory. On the other hand, if empirical SADs are unremarkable compared to their feasible sets, we may not be able to detect a biological signal in the SAD itself.

Locey and White documented the general form of the feasible set for combinations of S and N, and confirmed that observed SADs appear to fall within this general envelope according to R2. This establishes the feasible set as a logical baseline against which to evaluate SADs, but does not capture fine distinction between observed SADs that fall extremely close to, or deviate from, the central tendencies of their feasible sets. Quantifying these deviations, and any variation in their magnitude and direction, may be key to extracting biological information from the SAD.

Here, we extend their efforts by pinpointing the shapes of observed SADs within the range of possible shapes from their feasible sets. We use a new algorithm for rapid unbiased sampling of the feasible set space, and a compilation of abundance datasets for X communities, to (1) intensively sample the range of possible forms for a vast collection of SADs and (2) evaluate whether ecological communities consistently deviate from their statistical baselines or simply track their most-likely forms. Consistent with Locey and White, we find that most feasible sets are overwhelmingly dominated by hollow curves. Within this general “hollow curve” form, we document appreciable variation in shape as captured by both skewness and Simpson’s evenness. We show that empirical SADs vary in their position relative to their feasible sets, and that exceptionally skewed and uneven SADs are disproportionately common in ecological communities. We suggest refocusing research on the SAD to identify the correlates and develop explanations for these common, but not ubiquitous, deviations.

**Methods**

*Sampling the feasible set*

Previous efforts to characterize feasible sets for the SAD via uniform sampling have been constrained by computational feasibility. We developed a new algorithm capable of rapidly drawing orders of magnitude more samples from communities with vastly more species and individuals than was previously practical.

The feasible set is defined as all possible unique unordered vectors of *S* integers that sum to *N*. Finding the size and elements of the feasible set is a version of constrained integer partitioning problem, which the number theorists are deeply into but have not solved concisely.

As S and N increase, the number of elements of the feasible set expands rapidly. Rather than enumerate all possible forms, we sample the distribution. It is therefore crucial that we sample the distribution at random.

[describe Hao’s algorithm]

*Data*

We used a compilation of datasets originally created by Baldridge, White, and Xiao. This database contains X SADs spanning trees, birds, mammals, and [misc abund taxa]. Following Locey and White and Xiao et al, we restrict our analyses to 1 year of BBS data. We developed our analytical pipeline using annual plant survey data from the Portal Project, but do not include those results in this analysis as those data are replicated across time but not across space.

We note that others have removed SADs for communities with fewer than 10 species on the grounds that it is challenging to fit a distribution to so few points. Small communities do indeed present challenges for the feasible set approach, but this does not emerge except for *extremely* small ones. We therefore included even species-poor communities in the initial analysis, and note that we have low confidence in results from communities for which 10,000 draws did not return at least X unique samples from the feasible set.

*Estimated species richness*

Any census of a community likely underestimates the true number of species present. To correct for potentially missed species, we also ran our analyses on SADs corrected with estimates of the true S based on [names of the richness estimators]. We reasoned that omitted species are likely to be rare species, and adjusted observed data by adding 1 individual of each species estimated to be missed. We ran analyses for both unadjusted and corrected datasets.

*Characterizing the shape of the SAD*

We used skewness (e1071::skewness) and Simson’s evenness (vegan::diversity) to characterize the shape of every SAD. These metrics do not correspond directly to one another. Skewness is a statistical descriptor that measures skewness. Simpson’s evenness is more common in the community ecology context.

For every observed SAD, we calculated skewness and evenness values for 2500 samples from the corresponding feasible set. We then compared the skewness and evenness of the observed SAD to the distribution of values from the feasible set. The raw value of skewness or evenness for an SAD is not informative on its own, because the range of probable values varies depending on S and N. Therefore, we focus on where an observed distribution’s value falls relative to the distribution of values of the samples from its feasible set. This percentile value ranges from 0 to 100. At random, we would expect a uniform distribution of percentile values (i.e. 10% of the time, we expect a percentile value less than 5 or greater than 95).

Our sampling algorithm is stable and documented as an R package on GitHub at [www.github.com/diazrenata/feasiblesads](http://www.github.com/diazrenata/feasiblesads). All data and code required to replicate our analyses is archived as a research compendium on GitHub at [www.github.com/diazrenata/scadsanalysis](http://www.github.com/diazrenata/scadsanalysis).

**Results**

*General form and range of variation of SADs*

* The minimum S and N required for an appreciably large (>2500 unique samples) feasible set varies with both S and N. N > 9 and S < 45 appears to be too small. For larger N, S as low as 3 appear sufficient.
* The vast majority of samples are hollow curves. Depending on S and N, these range in evenness and skewness.
* There is more variation in skewness than in evenness.

*Position of empirical SADs within feasible set*

* Empirical SADs are often, but far from always, more skewed and less even than the bulk of their feasible sets. The means for the datasets are [][I bet in the 60-80th percentile range, *usually*], but there is substantial variation. Highly skewed and less even values are overrepresented compared to a random expectation, but far from a rule.
* The percentile value of the observed SAD relative to its feasible set is not strongly correlated with S, N, N/S, or the range of metric values represented in the feasible set.

*Impact of missed species*

* Adjusting communities using species richness estimators consistently yields *more* extreme values for both skewness and evenness.

**Discussion**

Many empirical SADs do appear to track their statistical constraints. However, in all datasets, highly skewed and uneven vectors are overrepresented. This is potentially a signal of some biological process that forces a more uneven distribution of abundance across species even than we would anticipate based on the feasible set.

Understanding the correlates and causes of the variation in the deviations between observed and random SADs will be a major new horizon. Are communities stable in the magnitude and direction of their deviation, or do they become more or less similar to their feasible sets over time? Do experimental manipulations or disturbances drive communities towards or away from the center of their feasible sets?

This method demonstrates a new and promising way to extract whatever information there is to be found in the SAD and leverage that to understand what makes some ecosystems deviate from a random baseline.

**Scrap**

Beyond the hollow curve

We have identified a new axis of variation for the SAD

* Species abundance distributions almost always follow a classic “hollow curve” pattern. There are a few abundant, and many rare, species, regardless of the system or taxon in question.
* Puzzlingly, most theories also predict hollow curves. This has made it difficult to use the species abundance distribution to evaluate competing theories.
* Recent work (Locey and White; Frank; Blonder?) has identified an underlying statistical constraint on the species abundance distribution.
* Locey and White found that most possible forms for the species abundance distribution are hollow curves. That is, in the absence of any strong force (or in the presence of numerous countervailing forces), we should in fact expect the species abundance distribution to follow some form of hollow curve based simply on the mathematical realities of partitioning N individuals across S categories.
* If this is the case, focusing on the “hollow curve” designation overlooks variation in the SAD that may hold the key to distinguishing between forms that are highly likely to emerge at random and those that are less likely. Even unlikely forms may be generally “hollow curves”; the important variation lies in the more nuanced aspects.
* Exposing variation in SADs in this regard may be the key to evaluating theoretical predictions for SADs and/or evaluating the possibility that, in aggregate, ecological communities behave essentially randomly.
* Previously, exploring the full range of SADS within the feasible set and comparing them to observed distributions has been prohibitively computationally intensive. Phenomenological vs stat mechanics framing...
* Phenomenological:
  + SADs are almost always hollow curves. This makes them hard to tell apart.
  + We can sample the feasible set to map the range of variation in possible forms, then compare a distribution of interest to this corpus
  + This allows us to differentiate between distributions that simply track their baseline and those that are unusual
* Statistical mechanics:
  + Either because there is no force operating on a system, or because numerous forces cancel each other out, it might appear random
  + Distinguishing between highly-likely-at-random and unlikely observations may help us identify the forces operating in ecological communities that shape drive SADs away from a random baseline.

Methods

* Sampling the feasible set
  + Define the problem
  + Explain algorithm
    - Gnomes:
      * Allocate 1 individual to all species.
      * Sequentially redistribute remaining individuals and species according to the possible ways.
      * (Hao wrote an actual robust version of this with 0 gnomes).
* Describing variation in distributions
  + We use skewness and Simpson’s diversity to summarize SADs.
  + Illustrate how Simpson’s and skewness map on to a FS
* Comparison to observed SADs