**Cover letter**

[Greeting]

We are writing to submit [ms title] as a Letter.

The species abundance distribution (SAD) is one of the most well-known and highly-studied patterns in ecology. However, efforts to understand it in biological terms have struggled because almost all SADs – including those generated by various ecological theories - conform to the same general form, with a few common species and many rare species. One explanation for the ubiquity of this pattern is that it is a statistical artefact of the mathematical process of dividing a given number of individuals into species. However, we may still be able to detect biological influences on this pattern, if the SADs we observe in nature deviate from the shapes that would emerge statistically. We use combinatorics to characterize the shapes that would emerge from statistical processes for the SAD for 24,500 ecological communities, and evaluate (1) whether observed SADs consistently deviate from these statistical baselines, and (2) how the specificity of the statistical baseline, and consequentially our statistical power to detect deviations between observations and the baseline, differs among very small and very large communities. We show that empirical SADs for large communities are consistently highly skewed and uneven relative to the SADs that emerge from strictly mathematical processes. However, we find that very small communities have considerably less narrowly defined statistical baselines, which greatly reduces our power to detect deviations between observed SADs and the baseline. This work highlights two novel avenues for understanding and interpreting the SAD, and ecological patterns in general. First, we are able to disentangle a biological signal from the apparent statistical constraint on the SAD, which will allow us evaluate theories in terms of how well they predict the *biological* signal in the SAD. Second, and more generally, this work directly engages with the question of to what extent we can leverage statistical baselines to interpret ecological patterns, and under what conditions this strategy breaks down.

This work goes beyond the efforts pioneered in Locey and White (2013) by leveraging the existence of the statistical constraint they document to specifically test for deviations, by explicitly exploring how the specificity of the statistical baseline, and correspondingly our ability to detect deviations, varies over large ranges in community size. We also substantially expand the range of community sizes and the number of empirical communities we analyze.

Thank you for your time and consideration.

Sincerely,

**Title: Empirical abundance distributions are more uneven than expected given their statistical baseline**

**Running title: Comparing SADs to their statistical baselines**

**Authors:** Renata M. Diaz, Hao Ye, S. K. Morgan Ernest

RMD: School of Natural Resources and Environment, University of Florida; corresponding author; renata.diaz@weecology.org

HY: University of Florida Health Science Center Libraries

SKME: Dept. of Wildlife Ecology and Conservation, University of Florida

**Authorship statement**

RMD and SKME conceived the analysis; HY devised the algorithm to sample the feasible set and reviewed the coded implementation; RMD conducted the analyses and wrote the first draft of the manuscript; all authors contributed substantively to revisions.

**Abstract**

Species abundance distributions consistently conform to a “hollow-curve” shape, with a large number of rare species and very few common species. One explanation for this is that the hollow-curve is an artefact of the statistical process of dividing a particular number of individuals into a given number of species. Whatever influences biological processes have on the shape of the species abundance distribution – and whatever leverage may exist for using the species abundance distribution to detect or diagnose such processes – may therefore be most detectable in deviations between observed abundance distributions and the shapes we would expect to emerge statistically. We compared the species abundance distributions for 24,500 empirically sampled communities to their corresponding statistical baselines, and found that empirical abundance distributions are consistently highly skewed and highly uneven relative to their statistical baselines. We also explored how the size of the community, in terms of the number of individuals and species, affects the specificity of the statistical baseline and consequentially our statistical power to detect deviations between observed distributions and the baseline. We identified an approximate size range below which we have considerably reduced sensitivity for detecting such deviations. Taken together, these results provide a new source of leverage for interpreting the species abundance distribution in biological terms, and help contextualize the promise and limits of using statistical baselines to study ecological pattern.

**Keywords**

Species abundance distributions; feasible set; combinatorics

**Introduction**

The distribution of how the total number of individuals in a community are partitioned among the species in that community, or 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 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 attempted to test numerous theories intended to determine which biological processes are most important for structuring assemblages of species by comparing theoretical predictions for the SAD to observed SADs (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 attention for both its consistent pattern and potential as a theoretical benchmark, the utility of the SAD for assessing the processes structuring ecological communities is currently unclear.

Accumulating evidence suggests that statistical constraints may actually generate the most striking feature of the species abundance distribution – the hollow curve (Locey and White 2013, White et al 2012, Frank 2019). Failing to account for these statistical considerations may have led us to focus on an obvious but relatively uninformative aspect of the SAD. Power-law or log-series distributions (i.e. ‘hollow-shaped’ curves) emerge as statistical attractors for abundance distributions generally (Frank 2009, 2019). In the specific case of the SAD, approaches using both statistical mechanics (i.e. the Maximum Entropy Theory of Ecology (METE); Harte et al 2008, Harte 2011) and combinatorics (i.e. ‘the feasible set’; Locey and White 2013) generate realistic hollow curves via the random division of the total number of individuals in a community, *N*,into the total number of species present, *S*. 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), 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, this does not necessarily mean that the SAD cannot be biologically informative. Biological factors may introduce relatively subtle, but meaningful, deviations between the shapes of observed SADs and the shapes of the SADs we would expect to observe simply due to the mathematical constraints imposed by *S* and *N,* which we refer to as the statistical baseline (Locey and White 2013, Harte and Newman 2014). It may be possible to use these *deviations* to detect strong ecological processes or evaluate theories (Harte and Newman 2014, Xiao et al 2016). If, as is often the case for large values of *S* and *N,* the vast majority of mathematically achievable SADs for a community share a similar shape (Locey and White 2013), an empirically observed SAD that deviates even slightly from this statistical baseline is unlikely to have emerged at random (Locey and White 2013). Such a deviation might be the signature of a non-statistical – i.e., biological – process operating on the relative abundances of species (Harte and Newman 2014). We can evaluate proposed processes or theories based on how well they predict these *deviations* between observed SADs and their statistical baselines, and not just the general, mathematically-forced shape for the distribution.

Successfully applying SADs in this fashion depends on our capacity to detect and quantify deviations between empirical observations and their statistical baselines, which requires metrics and computational approaches that allow us to quantify and interpret whatever deviations may exist. Here, we build upon the combinatoric approach developed by Locey and White (2013) to define and explore the statistical baselines for SADs given *S* and *N*. For a given *S* and *N*, there exists a finite (but potentially very large) set of possibledistributions of individuals into species – the *feasible set*. The distribution of shapes represented among the elements of the feasible set constitutes a statistical baseline for the SAD. If an observed SAD is simply drawn at random from the set of mathematically possible SADs, it is likely to have a shape similar to the shapes most common in the feasible set. The feasible set can therefore be used to compare observations to a statistical baseline, and to explore how the statistical baseline varies over gradients of S and N (Locey and White 2013).

It is especially important to understand how the statistical baseline varies with S and N, because our capacity to detect deviations between an observation and its statistical baseline is likely to depend on the size of the community in terms of *S* and *N,* with reduced sensitivity for very small communities. If S and N are small, there may be too fewpossible SADs in the feasible set to define a clear statistical baseline. When this occurs, we have reduced confidence that even an observation that deviates slightly from the statistical baseline could not reasonably have emerged at random from the relatively restricted pool of possible outcomes (Jaynes 1957). This general concern has been widely acknowledged as relevant to efforts to compare ecological observations to statistical baselines (Harte 2011, Locey and White 2013, White et al 2012), but there has not yet been a specific effort to quantify these effects for the SAD or to identify the range of community sizes most strongly affected. Because ecologists study the SAD for communities varying in size from the very small – *S* and *N* < 5 – to the enormous – *S* and *N* in the thousands and tens of thousands – it is important to identify the community sizes for which we can and cannot confidently detect deviations from the statistical baseline and contextualize our interpretations accordingly.

Here we use the feasible set to define statistical baselines for empirical SADs for 24,500 communities of birds, mammals, trees, and miscellaneous other taxa based on their 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 baseline, 2) how the specificity of the statistical baseline varies over gradients of *S* and *N*, and 3) whether this variation appears to be associated with variation in our capacity to detect deviations between observations and the corresponding baselines.

**Methods**

*Datasets*

We used 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 2016, Baldridge 2015). This compilation consists of cleaned and summarized community abundance data for trees obtained from the Forest Inventory and Analysis (Woudenberg et a al 2020) and Gentry transects (Phillipes and Miller 2002), birds from the North American Breeding Bird Survey (Sauer et al 2013), mammals from the Mammal Community Abundance Database (Thibault et al 2011), and a variety of less commonly sampled taxa from the Miscellaneous Abundance Database (Baldridge 2015). Because characterizing the random expectation for the SAD becomes computationally intractable for very large communities, and trivially uninformative for very small ones, we filtered our datasets to remove communities with more than 40720 individuals. This resulted in the removal of 4 communities from the Miscellaneous Abundance Database. 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 (2016) and analyzed only a single, randomly selected, year of data, because samples taken from a single community at different time points are likely to covary highly. 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. Prior to aggregating results across communities and datasets, we removed communities with only one species, or for which N = S, because these communities have only one possible SAD. After filtering exceptionally large or small communities and multiple years of data collection, our final dataset consisted of 24,519 communities with S and N ranging from 2 to 250 and 3 to 40714, respectively ([Figure](#_Figure_1:_Communities)). Details of the filtering process can be found at https://github.com/diazrenata/scadsanalysis/blob/clean-and-tests/analysis/reports/filtering.md.

*Generating the statistical baseline*

We use the concept of the *feasible set* to characterize a 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 complete set of these unique partitions is then the feasible set. Neither species nor individuals are distinguishable from each other; thus, partitions are unique if and only if they differ in the number of species that have a particular abundance (Locey and White, 2013). 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 because they contain they each contain one species with an abundance 1, 3, and 5, respectively, and they differ only in the *order* of the numbers. Alternative formulations of the SAD with different 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 this simple set of assumptions that has previously been shown to generate realistic statistical baselines (Locey and White 2013).

Characterizing the statistical properties feasible set can be computationally intensive, particularly for large combinations of S and N. While it is possible to list all possible partitions for small *S* and *N*, the number of elements in the feasible set increases rapidly with S and N. This renders it necessary to draw samples from the feasible set, rather than enumerating all of its elements. Unbiased sampling of large feasible sets is a nontrivial computational problem that 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 drew 4000 samples from the feasible set to characterize the distribution of statistically probable shapes for the SAD. We filtered the 4000 samples to unique elements. For small values of S and N, it can be impossible or highly improbable for the 4000 samples from the feasible set to all be unique, but for large communities, all 4000 are usually unique. We refer to this as the sampled feasible set.

*Comparing observed SADs to their baselines*

In the absence of any other process, an SAD with a particular S and N will reflect the statistical properties of the feasible set for the same S and N. We focus on the shape of the distribution of abundances across species as the characteristic of interest. Metrics related to this shape 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 measures the asymmetry of a distribution around its mean, and Simpson’s evenness is a commonly used metric in ecology for assessing how equitably abundance is distributed across species. By calculating these metrics for each of the samples in the community’s sampled feasible set (see *Generating the statistical baseline* above), we generated a distribution describing the general shape (i.e. evenness or skewness) that we expect from the statistically-generated SADs. Note that skewness, as implemented in the R package “e1071” (Meyer et al 2019), always evaluates to 0 for distributions with only two species and cannot be computed for distributions where all values are equal, and we excluded those cases from analyses of skewness.

To assess whether the shape of an observed SAD was statistically unlikely, we calculated Simpson’s evenness and skewness for the observed SAD and compared these observed values to the distributions of evenness and skewness obtained from that community’s sampled feasible set. An observed SAD’s deviation from its feasible set was determined by computing the percentile rank of its skewness and evenness relative to the sampled distributions for skewness and evenness, respectively. These percentile ranks are comparable across different community sizes, allowing broad-scale assessment across wide ranges of S and N. After aggregating across communities, if community abundance distributions reflect random draws from their feasible sets, their percentile rank values should 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. Because an earlier survey in this space (Locey and White 2013) found that the tendency is for empirical SAD to be more skewed and less even than their feasible sets, we used one-tailed 95% confidence intervals and tested for unusually *high* values for skewness and *low* values for evenness. This comparison is not meaningful if there are very few unique values in the distributions of skewness and evenness, which can occur for small feasible sets. We therefore excluded communities for which the distribution of skewness or evenness values from the sampled feasible set had fewer than 20 unique values (in these cases, it is impossible for an observation to fall above or below the 95th or 5th percentile, respectively).

*The narrowness of the expectation*

We also used the distributions of skewness and evenness from the sampled feasible set to describe the relative specificity of the statistical baseline, in order to assess whether there could be challenges in determining whether observed communities differ from their statistical baselines. We quantified the narrowness of a distribution as the ratio of the range of values encompassed within a 95% density interval relative 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](https://github.com/diazrenata/scadsanalysis/blob/clean-and-tests/analysis/reports/self_similarity.md)).

**Results**

*Observed SADs 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:_Overall), [Table](#_Table_1:_Percentile) ). 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. These outcomes contrast with the results from the FIA dataset, for which percentile scores were near-uniformly distributed for skewness (5% of observations are more skewed than 95% of the feasible set), and less concentrated at the extreme than any of the other datasets for evenness (9% of observations are less even).

*The narrowness of the expectation*

Across the communities we analyzed, the statistical baseline for both skewness and evenness narrows considerably as the size of the feasible set increases ([Figure](#_Figure_2:_95%), [Figure](#_Figure_5:_95%)). The size (or number of elements) of the feasible set increases with species richness (S), total abundance (N), and average abundance (N/S). At the extreme, for communities with relatively small feasible sets – fewer than approximately 1000 elements for skewness, and approximately 200 elements for evenness – the range of the 95% interval of values in the distribution approaches the entire range of values. Among the datasets we analyzed, the FIA database is the most dominated by communities for which the 95% interval spans nearly the entire range of values (for skewness), or a large proportion of the range (for evenness), reflecting relatively broad and nonspecific statistical expectations for the shape of the feasible set ([Figure](#_Figure_7:_Distribution)).

**Discussion**

We found widespread evidence that the shapes of the SADs for a range of real ecological communities are more skewed and less even than we would expect given their feasible sets. These deviations most likely signal that ecological processes operate on top of statistical constraints and drive the SAD away from the shape it would assume in the absence of a dominating non-statistical process. Our results suggest that the prevailing processes structuring these communities cause abundance distributions to be more uneven – rather than those that cause individuals to be spread evenly across species. These processes might be those that promote the persistence of rare species at extremely low abundances (e.g. Yenni et al 2012) –thereby lengthening the rare tail of the SAD – or processes that encourage or allow dominant species to be hyper-dominant without driving other species entirely to extinction (Chesson 2000). Although a disproportionate number of communities deviated statistically their feasible sets, there were also many comparable communities for which we did not detect deviations. In such cases, numerous ecological processes may operate simultaneously and with countervailing impacts on abundance distributions, resulting in no dominating net effect on the shape of the distribution beyond that imposed by fundamental constraints (Harte 2008; Harte and Newman 2014). Going forward, testing whether ecological theories or common functional approximations accurately predict this range of variation in deviations between observed SADs and their expectations may be much more fruitful than focusing only on the general form of the SAD, which may emerge from statistical constraints (McGill et al 2007; Locey and White 2013).

Unlike the other four datasets, communities in the FIA dataset showed weak or no evidence of deviations from their feasible sets. These results may be an artifact of statistical issues related to community size. The FIA communities are by far the smallest across our datasets. Communities with small values of S and N have smaller feasible sets, which affects our ability to detect deviations from the most probable shapes for the SAD. When there are relatively few possible SADs, the distributions of evenness and skewness values derived from the feasible set are less narrowly peaked, meaning there is a relatively weak statistical distinction between “common” and “extreme” shapes for the SAD. When we compared the distributions of shape metrics for small communities to those for large ones, we found that small communities generate broader distributions of evenness, and especially skewness, than those for large communities (Figure). For such communities, the deviations – or lack thereof – that we perceive are less informative than for larger communities with more strongly defined statistical baselines (Jaynes 1957).

Additionally, if the lack of discernable deviations from the feasible set for the FIA communities is indeed a byproduct of their generally small size, then we would expect similarly-sized communities from other datasets to have similar results. We identified 371 communities from other datasets with S and N matching communities in the FIA, and found no difference in the distribution of percentile scores between communities from FIA and communities from other datasets ([Figure](#_Figure_8:_Direct), [Table](#_Table_2:_Percentile)). Although this is a highly restricted subset of communities relative to the 20,000 FIA communities we analyzed, these results point to community size, and not biological features specific to FIA, as a likely explanation for the weak evidence for deviations across the full FIA dataset.

If this is indeed the case, it means that small-community considerations may affect our capacity to meaningfully distinguish signal from randomness using this approach. FIA communities, with their broad distributions of shape metrics and overall lack of detectable signal, have on the order of 10 species and 50-100 individuals. While these values do not constitute hard thresholds, they may indicate a general range of values below which we have relatively diminished power to detect deviations from the statistical baseline represented by the feasible set. To meaningfully draw inferences from deviations in these small communities, we will likely need more sensitive metrics (than skewness and evenness), and/or theories with stronger assumptions on the SAD to make comparisons against. In the absence of such, we may stand to learn the most by focusing on SADs from relatively large communities.

It is also important to recognize that there are multiple plausible approaches to defining a statistical baseline for the SAD, of which we have taken only one (Locey and White 2013; Haegeman and Loreau 2008). Our approach follows Locey and White (2013) and reflects the random partitioning of individuals into species, with the resulting distributions considered unique if the species’ abundance values are unique, regardless of the order in which the values occur. Biologically, differences in *order* would correspond to differences in *which* species contain the most or least individuals. This philosophy reflects a longstanding approach in the study of abundance distributions; that is, to focus on the shape of the distribution without regard to species’ identities (McGill 2007). However, there has yet to be a direct examination of either the validity or outcomes of alternative methods for generating statistical baselines for the SAD. Other formulations for the statistical baseline may be equally valid and generate different statistical expectations, including forms that approximate exponential, Poisson, or log-series distributions (Favretti 2018, Harte et al 2008). Comparing the results that emerge from different baselines will be an important next step towards reinvigorating the use of the SAD as a diagnostic tool.

Our study demonstrates both the utility, and the potential challenges, of applying tools from the study of complex systems and statistical mechanics to study ecological communities (Harte 2008, Harte and Newman 2014, Haegeman and Loreau 2008, White et al 2012)**.** While concepts such as maximum entropy and the feasible set are promising new horizons for macroecology, the small size of some ecological communities may present difficulties that do not occur as often in the domains for which these tools were originally developed (Haegeman and Loreau 2008, Jaynes 1957). If a substantial contingent of ecological communities are too small to generate highly resolved statistical baselines, these approaches may be less informative than we might hope – as appears to be the case for the smallest communities in our analysis. However, our application of these approaches in larger communities demonstrates that while mathematical constraints appear to have strong effects on the general form of the SAD, these constraints alone do not fully account for the extremely uneven SADs we often observe in nature – leaving an important role for ecological processes. This ability to detect and diagnose the specific ways in which empirical SADs deviate from randomness can open up new avenues for understanding how and when biological drivers affect is shape. There is, of course, still many facets to be improved in our ability to distinguish biological signal from randomness, including assessing alternative statistical baselines and calibrating our expected power to detect deviations, especially for small communities. Indeed, more sensitive metrics could also enable identification of processes that operate through time (note that, in this analysis, we sampled each community at a single point in time). 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

##### 

Distribution of communities from each dataset in terms of their total abundance (N) and species richness (S).

##### Figure 2: 95% ratio illustration

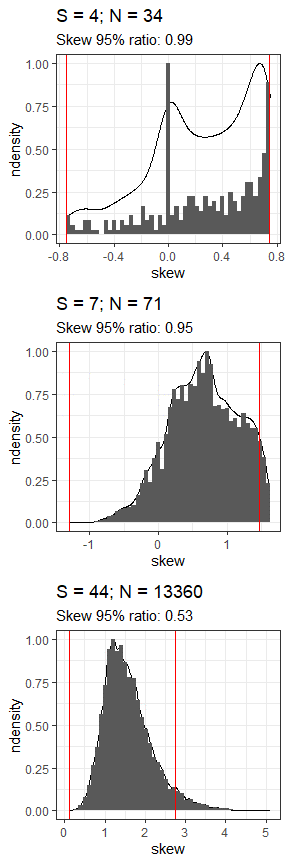
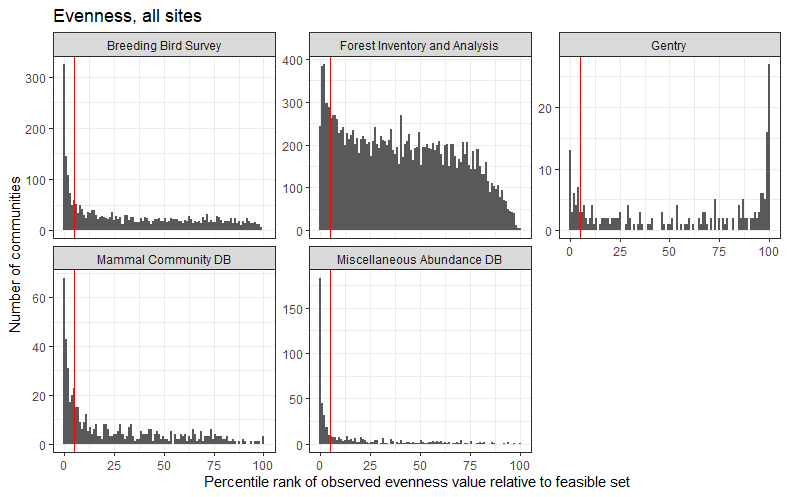


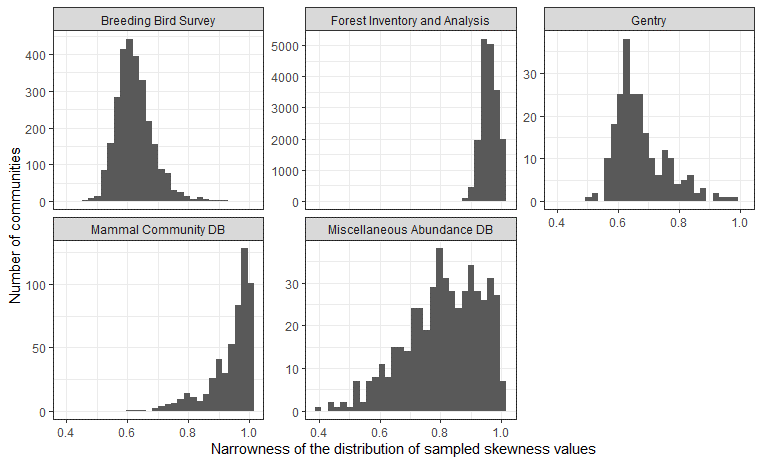
Illustration of the relationship between the feasible set and the statsitical expectation for hypothetical example communities with S = 4, 7, or 44 and N = 34, 71, or 13360 (top to bottom). Samples of unique SADs are drawn from a community’s feasible set. Feasible sets can range substantially in size, depending on the community size (e.g., a feasible set of 297 unique distributions for S=4 and N=34 vs 6.5e+70 unique distributions for S=44, N=13360). For every SAD drawn from the feasible set (left column), we calculate the skewness (color scale) or evenness (not shown). These values generate the distribution of expected statistic values from the feasible set (right column). The ratio of the range encompassed in the one-tailed 95% interval (space between red lines, right), compared to the full range of values for the statistic, describes how narrowly peaked or broad the distribution is. This ratio tends to decrease as the size of the feasible set increases and the distribution becomes more narrowly defined (top to bottom).

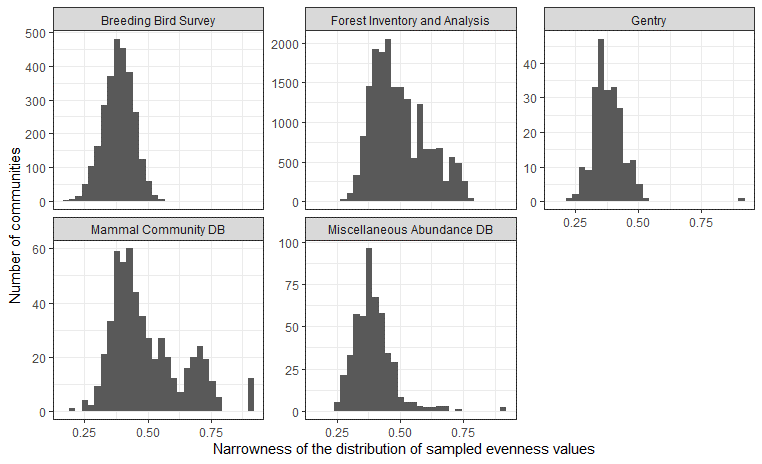
##### Figure 3: Overall percentile results



Histograms of percentile ranks for observed values of skewness (top) and evenness (bottom) relative to the distributions of values from the sampled feasible set for all communities. These plots exclude communities with fewer than 20 unique values for skewness or evenness in the sampled feasible set, and plots for skewness exclude communities with fewer than 3 species. The vertical red line marks the 95th percentile for skewness and the 5th percentile for evenness. At random, percentile ranks should be uniformly distributed from 0 to 100, and no more than 5% of values should be above or below the 95th and 5th percentiles, respectively.

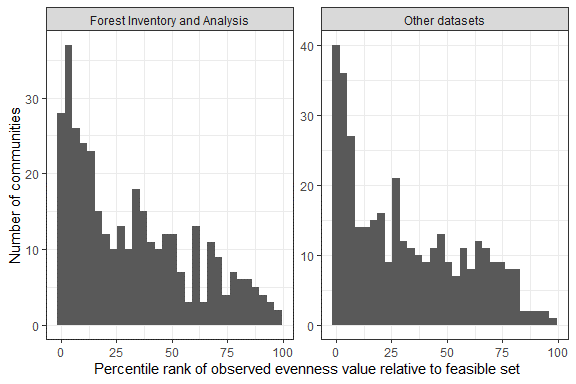
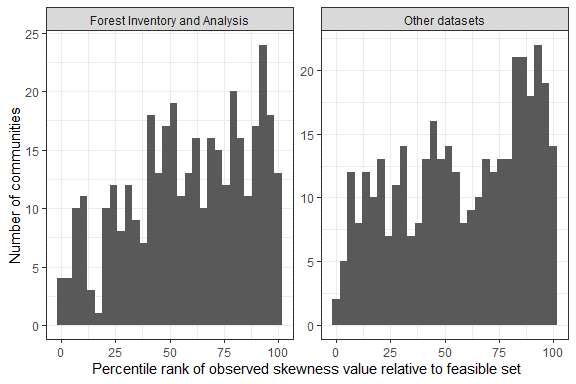
##### Figure 4: Distribution of 95% intervals by dataset





Histograms of the narrowness of the distribution of skewness (top) and evenness (bottom) values for the sampled feasible sets for communities from each dataset, excluding communities with fewer than 20 unique values for skewness or evenness and, for skewness, fewer than 3 species. The narrowness of the distribution is described as the ratio of the range of the one-tailed 95% interval to the full range of values. This value ranges from 0-1, with higher values indicating broader distributions.

##### Figure 5: Direct comparison of FIA and similarly sized sites



Histograms of percentile ranks for observed values of skewness (left) and evenness (right) relative to the distributions of values from the sampled feasible set for 371 pairs comprising a FIA community and a community from another dataset with the same S and N as the FIA community. These plots exclude communities with fewer than 20 unique values for skewness or evenness in the sampled feasible set. At random, percentile ranks should be uniformly distributed from 0 to 100.

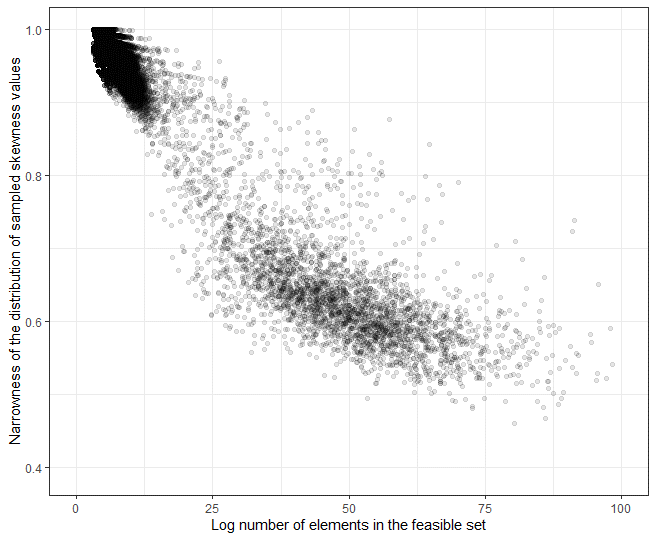
**Supplement Figures & Tables**

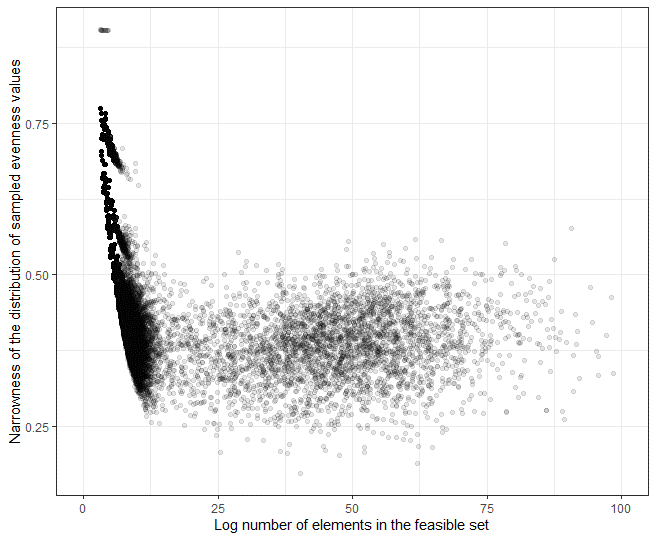
##### Table 1: Percentile results (**move to supplement)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dataset** | **Proportion of communities with skewness above 95th percentile** | **Number of communities analyzed for skewness** | **Proportion of communities with evenness below 5th percentile** | **Number of communities analyzed for evenness** |
| Breeding Bird Survey | 0.130184 | 2773 | 0.259647 | 2773 |
| Forest Inventory and Analysis | 0.054208 | 18300 | 0.093966 | 18113 |
| Gentry | 0.188341 | 223 | 0.151786 | 224 |
| Mammal Community DB | 0.158287 | 537 | 0.354244 | 542 |
| Miscellaneous Abundance DB | 0.345529 | 492 | 0.595918 | 490 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dataset** | **Proportion of communities with skewness above 95th percentile** | **Number of communities analyzed for skewness** | **Proportion of communities with evenness below 5th percentile** | **Number of communities analyzed for evenness** |
| Forest Inventory and Analysis | 0.054208 | 18300 | 0.093966 | 18113 |
| Other datasets | 0.163478 | 4025 | 0.307272 | 4029 |

##### Figure 5: 95% intervals vs. size of feasible set **(move to supplement)**





The narrowness of the distribution of skewness (top) and evenness (bottom) values from the sampled feasible set compared to the number of elements in the feasible set. The narrowness of the distribution is described as the ratio of the range of the 0-95th quantile values (for skewness), or the 5th-100th quantile values (for evenness), compared to the range of values from the entire sample. This ratio ranges from 0-1, with larger values indicating broader distributions and less-specific statistical expectations. The number of elements in the feasible set increases with increasing S, N, and average abundance (N/S). Both plots exclude feasible sets with fewer than 20 unique values for skewness or evenness, and skewness also excludes communities with S < 3.

##### Table 2: Percentile results comparing FIA to direct counterparts **(move to supplement)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dataset** | **Proportion of communities with skewness above 95th percentile** | **Number of communities analyzed for skewness** | **Proportion of communities with evenness below 5th percentile** | **Number of communities analyzed for evenness** |
| Forest Inventory and Analysis | 0.0808625 | 371 | 0.1780822 | 365 |
| Other datasets | 0.0862534 | 371 | 0.2109589 | 365 |

**References**

Baldridge, E. (2015). *MiscAbundanceDB\_main*. <https://doi.org/10.6084/m9.figshare.95843.v4>

Baldridge, E., Harris, D. J., Xiao, X., & White, E. P. (2016). An extensive comparison of species-abundance distribution models. *PeerJ*, *4*, e2823. <https://doi.org/10.7717/peerj.2823>

Chesson, P. (2000). Mechanisms of Maintenance of Species Diversity. *Annual Review of Ecology and Systematics*, *31*(1), 343–366. <https://doi.org/10.1146/annurev.ecolsys.31.1.343>

Favretti, M. (2018). Remarks on the Maximum Entropy Principle with Application to the Maximum Entropy Theory of Ecology. *Entropy*, *20*(1), 11. <https://doi.org/10.3390/e20010011>

Fisher, R. A., Corbet, A. S., & Williams, C. B. (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*(1), 42–58. JSTOR. <https://doi.org/10.2307/1411>

Frank, S. A. (2009). The common patterns of nature. *Journal of Evolutionary Biology*, *22*(8), 1563–1585. <https://doi.org/10.1111/j.1420-9101.2009.01775.x>

Frank, Steven A. (2019). The common patterns of abundance: The log series and Zipf’s law. *F1000Research*, *8*, 334. <https://doi.org/10.12688/f1000research.18681.1>

Haegeman, B., & Loreau, M. (2008). Limitations of entropy maximization in ecology. *Oikos*, *117*(11), 1700–1710. <https://doi.org/10.1111/j.1600-0706.2008.16539.x>

Harte, J., Zillio, T., Conlisk, E., & Smith, A. B. (2008). Maximum Entropy and the State-Variable Approach to Macroecology. *Ecology*, *89*(10), 2700–2711. <https://doi.org/10.1890/07-1369.1>

Harte, John. (2011). *Maximum Entropy and Ecology: A Theory of Abundance, Distribution, and Energetics*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199593415.001.0001>

Harte, John, & Newman, E. A. (2014). Maximum information entropy: A foundation for ecological theory. *Trends in Ecology & Evolution*, *29*(7), 384–389. <https://doi.org/10.1016/j.tree.2014.04.009>

Jaynes, E. T. (1957). Information Theory and Statistical Mechanics. *Physical Review*, *106*(4), 620–630. <https://doi.org/10.1103/PhysRev.106.620>

Lawton, J. H. (1999). Are There General Laws in Ecology? *Oikos*, *84*(2), 177. <https://doi.org/10.2307/3546712>

Locey, K. J., & White, E. P. (2013). How species richness and total abundance constrain the distribution of abundance. *Ecology Letters*, *16*(9), 1177–1185. <https://doi.org/10.1111/ele.12154>

McGill, B. J., Etienne, R. S., Gray, J. S., Alonso, D., Anderson, M. J., Benecha, H. K., Dornelas, M., Enquist, B. J., Green, J. L., He, F., Hurlbert, A. H., Magurran, A. E., Marquet, P. A., Maurer, B. A., Ostling, A., Soykan, C. U., Ugland, K. I., & White, E. P. (2007). Species abundance distributions: Moving beyond single prediction theories to integration within an ecological framework. *Ecology Letters*, *10*(10), 995–1015. <https://doi.org/10.1111/j.1461-0248.2007.01094.x>

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2019). *E1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. <https://CRAN.R-project.org/package=e1071>

Phillips, O., & Miller, J. S. (2002). *Global patterns of plant diversity: Alwyn H. Gentry’s forest transect data set* (Vol. 89). Missouri Botanical Press.

Sauer, J. R., Link, W. A., Fallon, J. E., Pardieck, K. L., & Ziolkowski, D. J. (2013). The North American Breeding Bird Survey 1966–2011: Summary Analysis and Species Accounts. *North American Fauna*, *79 (79)*, 1–32. <https://doi.org/10.3996/nafa.79.0001>

Supp, S. R., & Ernest, S. K. M. (2014). Species-level and community-level responses to disturbance: A cross-community analysis. *Ecology*, *95*(7), 1717–1723. <https://doi.org/10.1890/13-2250.1>

Thibault, K. M., Supp, S. R., Giffin, M., White, E. P., & Ernest, S. K. M. (2011). Species composition and abundance of mammalian communities. *Ecology*, *92*(12), 2316–2316. <https://doi.org/10.1890/11-0262.1>

White, E. P., Thibault, K. M., & Xiao, X. (2012). Characterizing species abundance distributions across taxa and ecosystems using a simple maximum entropy model. *Ecology*, *93*(8), 1772–1778. <https://doi.org/10.1890/11-2177.1>

Woudenberg, S. W., Conkling, B. L., O’Connell, B. M., LaPoint, E. B., Turner, J. A., & Waddell, K. L. (2010). The Forest Inventory and Analysis Database: Database description and users manual version 4.0 for Phase 2. *Gen. Tech. Rep. RMRS-GTR-245. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 336 p.*, *245*. <https://doi.org/10.2737/RMRS-GTR-245>

Xiao, X., O’Dwyer, J. P., & White, E. P. (2016). Comparing process-based and constraint-based approaches for modeling macroecological patterns. *Ecology*, *97*(5), 1228–1238. <https://doi.org/10.1890/15-0962.1>

Yenni, G., Adler, P. B., & Ernest, S. K. M. (2012). Strong self-limitation promotes the persistence of rare species. *Ecology*, *93*(3), 456–461. <https://doi.org/10.1890/11-1087.1>