**Introduction**

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

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

If SADs are statistically inclined to be hollow curves without requiring biological influences, it is no surprise that we have struggled to interpret the hollow curve in biological terms. However, just because the main feature of the SAD may be statistically determined, it does not necessarily follow that the SAD cannot be biologically informative. Biological processes may cause relatively subtle, but meaningful, deviations between observed SADs and their statistical expectations (Locey and White 2013, Harte and Newman 2014). We may be able to use those *deviations* to evaluate theories (Harte and Newman 2014, Xiao et al 2016) by interpreting SADs not in terms of their absolute shape, but in terms of their shape relative to what we would expect simply due to statistical processes operating within the constraints of S and N(Frank 2009, Locey and White 2013). Using SADs in this fashion depends on our capacity to detect and quantify deviations between empirical observations and randomness, which requires metrics and computational approaches that allow us to quantify and interpret whatever deviations may exist, across many different types of ecological communities.

In particular, our ability to detect deviations may depend strongly on the size of the community (i.e. its values of S and N), because the statistically-derived expectation for the shape of the SAD may be more well-resolved, and thus a more powerful tool for inference, for communities with larger values of S and N. The underlying logic echoes a common theme in statistical mechanics and related fields, where the objective is to generate an expectation for the large-scale characteristics of systems that are made up of numerous subcomponents that can fall into different arrangements at the small scale (Haegeman and Loreau 2008). For a statistical expectation for an SAD with N individuals and S species, the arrangements are the possible divisions of N individuals into S species, and the large-scale characteristic of interest is the shape of the resulting SAD. (as captured by metrics such as skewness and evenness). By calculating the shape metrics for the range of *possible* divisions of individuals into species, we obtain a distribution of values for these metrics that reflects the values we would expect to observe if we were to randomly select a division at random from the array of possible divisions. In general, when a system involves very large numbers of subcomponents and therefore has numerous possible arrangements, nearly all of the arrangements tend to cluster around one set of large-scale characteristics (Jaynes 1979, Haegeman and Loreau 2008). In these scenarios, an observation whose large-scale characteristics deviate even slightly from those shared by the majority of possible arrangements is an indication that the information and assumptions that generated the expectation are incorrect or at least missing an important process at play in the system (Jaynes 1979). For the SAD, this would manifest as a narrow, peaked distribution of values for the shape metrics across the array of possible divisions. If the observed SAD deviates from this expectation derived from only S and N, it suggests that S and N alone do not provide enough information to predict the shape - perhaps because other processes are driving the system away from what would occur at random given S and N. Crucially, expectations obtained in this way are most informative when most of the possible small-scale arrangements appear similar at large scale – i.e., when the distribution of expected outcomes is very narrowly peaked (Jaynes 1979, Haegeman and Loreau 2008). If the array of possible arrangements encompasses broader and more even variation in large-scale characteristics, the expectation is less well-resolved and less informative (Jaynes 1979). When this occurs, we have less confidence that an observation that differs from the *most* likely characteristics is inconsistent with the expectation, because the expectation itself is nonspecific (Jaynes 1979, Haegeman and Loreau 2008). Small systems may be more likely to generate relatively broad expectations, because they have fewer subcomponents and fewer possible arrangements (Haegeman and Loreau 2008). This may be particularly relevant to ecology, because ecological systems can have considerably fewer subcomponents than is usual for the systems usually studied via statistical mechanics (Haegeman and Loreau 2008). However, it is not clear whether ecological communities are affected by these small-size phenomena, and if so, which community sizes (in terms of S and N) are affected the most.

Here we build upon the combinatoric approach developed by Locey and White (2013) to establish if, and in what ways, empirical SADs deviate from a statistical expectation based on S and N. We also examine whether smaller communities indeed have less narrowly-defined statistical baselines than large ones, and whether this appears to modulate our capacity to identify deviations. For [# communities] communities spanning X Y and Z taxa, we use combinatorics to generate the distribution of possible states for the SAD – i.e., a statistical baseline – for each community based on its number of species and number of individuals. We then compare *observed* SADs to their corresponding statistical baselines and evaluate 1) if the shapes of observed SADs are consistently unusual given their statistical expectations, 2) how the narrowness of the statistical expectation varies over gradients of *S* and *N*, and 3) whether this variation is associated with variation in whether observations can be distinguished from the expectation.

**Methods**

*Datasets*

We use a compilation of community abundance data for trees, birds, mammals, and miscellaneous other taxa that has been used in recent macroecological explorations of the SAD (White et al 2012, Baldridge et al 2014). This compilation consists of cleaned and summarized community abundance data for trees obtained from the Forest Inventory and Analysis (ref) and Gentry plots (ref), birds from the North American Breeding Bird Survey (ref), mammals from the Mammal Community Abundance Database (ref), and a variety of less commonly sampled taxa from the Miscellaneous Abundance Database (ref). We filtered these data to remove exceptionally large or small communities, because characterizing the random expectation for the SAD becomes computationally intractable for very large communities, and becomes trivially uninformative for very small ones. We therefore filtered our datasets to remove communities with more than X species or X individuals, or fewer than 2 species or X individuals. We also removed communities for which N = S, because these communities have only one possible SAD. We further filtered the FIA database, which contains roughly 100,000 communities of which approximately 90,000 have fewer than 10 species. Rather than analyze all of these small communities, we randomly selected 10,000 small communities to include in the analysis. All FIA communities with more than 10 species were included in the analysis. Finally, for sites that had repeated sampling over time, we followed White et al (2012) and Baldridge et al (2014) and analyzed only a single, randomly selected, year of data. It should be noted that our analyses include data from the Mammal Community Database and Miscellaneous Abundance Database that were collected over longer timescales and cannot be disaggregated, with an average temporal scale of X. After filtering exceptionally large or small communities and multiple years of data collection, our final dataset consisted of X communities encompassing X taxa, with S and N ranging from 2 X and X to X, respectively ([Figure](#_Figure_1:_Communities)). ). Code detailing the filtering process can be found at X.

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

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

For every community in our database, we uniformly drew 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 to randomly draw 4000 unique samples from the feasible set, 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 is likely to reflect whatever characteristics are common among the elements of its feasible set. We focu on the shape of the distribution of abundances across species as the characteristic of interest. Metrics related to this shape are frequently used in the study of community structure and have been used specifically in the context of distinguishing observed SADs from the feasible set (Locey and White 2013). We focus on two metrics to describe the shape of the SAD, skewness and Simpson’s evenness. Skewness measures the asymmetry of a distribution around its mean, with a negative skew indicating a longer tail of values to the left of the mean and a positive skew indicating the converse. 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 unique 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 randomly generated SADs. Note that skewness [breaks under specific circumstances – s < 3, all abundances equal], and we exclude 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 assessing what percent of values in its sampled distributions for skewness and evenness were below the observed values. This percentile rank is comparable across different community sizes, allowing broad-scale assessment across wide ranges of S and N. When we compare across communities, if communities tend to reflect random draws from their feasible set, we would expect their percentile rank values to be uniformly distributed from 0 to 100. However, if observed SADs are consistently more skewed or even than their feasible sets, the percentile values will be disproportionately concentrated towards the extremes. This logic breaks down if there are very few unique values in the distribution of possible values, which can occur for small feasible sets. We excluded instances for which the distribution of possible values had fewer than 20 unique values, because in these cases, it is impossible for an observation to fall above or below the 95th or 5th percentile.

*The narrowness of the expectation*

We also used these distributions skewness and evenness from the sampled feasible set to describe the relative narrowness or breadth of the statistical expectation over gradients of S and N, in order to assess whether there could be challenges in determining whether small communities differ from their randomly-generated expectations. We quantified the narrowness of a distribution as the ratio of the range of values encompassed within a 95% interval to the full range of values in the distribution ([Figure](#_Figure_2:_95%)). This allowed us to isolate and compare the *narrowness* of distributions that are not necessarily normally distributed and vary considerably in their actual values, means, ranges, and standard deviations. This metric corresponds qualitatively to more computationally-intensive approaches to measuring the self-similarity of the elements of feasible sets (Supplement).

**Results and Discussion**

We found widespread evidence that the shapes of the SADs for a range of real ecological communities are statistically unlikely when compared to their feasible sets. For four of the five datasets we analyzed – BBS, Gentry, Mammal Communities, and Misc. Abund – empirical SADs are highly skewed and highly uneven relative to their feasible sets much more frequently than would be expected by chance ([Figure](#_Figure_3:_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. However, we detected considerably less pronounced deviations for communities from the FIA communities ([Figure](#_Figure_3:_Overall), [Table](#_Table_1:_Percentile)), for which 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 (9% of observations are less even). Why the FIA communities show a different pattern is unclear, but may be partially due to their small size, which we explore further below. For the four other datasets, however, our results suggest that nonrandom processes drive observed SADs to be less even than would occur by chance.

These widespread and consistent deviations between empirical SADs and their feasible sets may be the signature of ecological processes operating on top of statistical constraints, and therefore offer new leverage for developing and evaluating theories to predict the shape of the SAD. One logical next step is to test whether existing theories (e.g. neutral theory, METE) and common functional approximations (logseries, exponential, and log normal) can accurately predict deviations from the feasible set. Many of these theoretical predictions generate hollow curves, but have yet to be evaluated on their ability to accurately capture nuanced deviations of the type we document here. In addition, our results suggest that the prevailing processes that structure communities tend to be ones that push abundance distributions towards an uneven state, rather than those that cause individuals to be spread evenly across species. These could be processes that promote the persistence of rare species at extremely low abundances – thereby lengthening the rare tail of the SAD – or processes that encourage dominant species to be hyper-dominant without driving other species entirely to extinction. Identifying the processes at play in particular systems will require further exploration, but the approach we have demonstrated here helps isolate the signal we can try and explain.

We found considerably weaker evidence for deviations for one of our five datasets, perhaps because of statistical issues related to community size. Unlike the other four datasets, communities in the FIA dataset showed at most weak evidence of deviations from their feasible sets. These communities are by far the smallest in our overall database: [most of them] have fewer than [X species/N individuals]; only [x%] of communities from the other four datasets are within this size range. For the subset of FIA communities with direct counterparts, in terms of S and N, in other datasets, we found no difference in the distribution of percentile scores between FIA and other datasets ([Figure](#_Figure_8:_Direct), [Table](#_Table_2:_Percentile)). Although this is a highly restricted subset of sites – approximately 370 points of comparison, relative to the 20,000 FIA sites we analyzed – the lack of a difference between FIA and other comparable sites points to community size, and not biological features specific to FIA, as a likely explanation for the relatively weak evidence for deviations across the full FIA dataset.

Community size may affect our ability to detect deviations via its potential effect on how similar the elements of the feasible set are to each other. Small communities may not have enough possible arrangements if their subcomponents, or elements in their feasible sets, to generate highly resolved distributions for the most probable shapes. When we compared the distributions of shape metrics for small communities to those for large ones, we found that samples from the feasible sets for small communities generate broader distributions of evenness, and especially skewness, than those for large communities ([Figure](#_Figure_5:_95%)). These broad distributions may not be specific enough to constitute a strong statistical expectation, and the deviations – or lack thereof – that we detect are correspondingly less informative. Compared to the other datasets, FIA is dominated by small communities with small feasible sets and broad distributions ([Figure](#_Figure_7:_Distribution)). This may explain why we do not detect strong differences between the observed SADs for FIA sites and the expectations we derive from their feasible sets.

If this is indeed the case, it means that small-community considerations may affect our capacity to meaningfully distinguish signal from randomness for communities with ranges of S and N that are quite common in ecology. As a group, the FIA communities had especially broad distributions of shape metrics, and exhibited an overall lack of detectable signal. Most of these communities 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 distinguish deviation from randomness. Unless we can develop more sensitive methods for identifying deviations even for these small communities, we may stand to learn the most by focusing on SADs from relatively large communities. In the meantime, sampling the range of forms represented in the feasible set helps us identify when the distribution of shapes present is relatively broad and therefore less informative.

This phenomenon also has more general implications for the trend in macroecology to use tools from the study of complex systems and statistical mechanics to study ecological communities (Harte….)**.** 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 challenges that do not occur as often in the large systems for which these tools were originally developed (Haegeman and Loreau). For example, maximum entropy and related approaches rely on there being a steeply-peaked distribution of probable outcomes for a system (H & L, Jaynes). If a substantial contingent of ecological systems have broad, nonspecific distributions of probable outcomes, these approaches may be less informative than we might hope. As we continue to explore the possible applications of the logic of probability and statistical mechanics to ecology, we should emphasize methods that allow us to detect and, if possible, adjust for the limitations imposed by small community size.

It is also important to recognize that there are multiple plausible approaches to the defining a statistical baseline for the SAD, of which we have taken only one (Locey and White 2013; Haegeman and Loreau 2008). Other formulations for the statistical baseline may be equally valid and can generate different statistical expectations, including forms that approximate exponential, Poisson, or log-series distributions. There is currently no unambiguous logical argument for one baseline over another, but comparing the results that emerge from different baselines is clearly an important next step in this process towards reinvigorating the use of the SAD as a diagnostic tool.

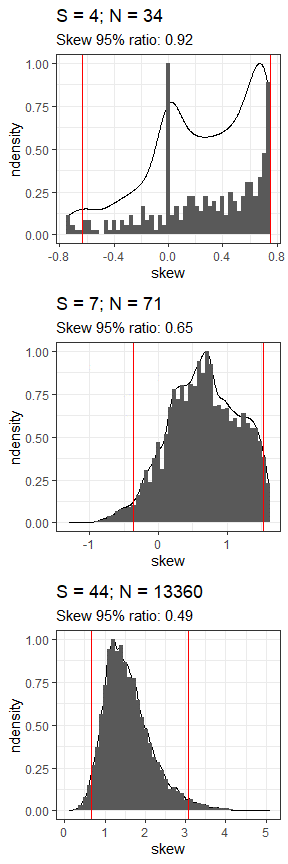
Characterizing and adjusting for a statistical baseline for the SAD, as we demonstrate here, refreshes our perspective on the distribution and opens up several new avenues for better understanding how and when biological drivers affect its shape. Persistent deviations between observed communities and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

**Figures**

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

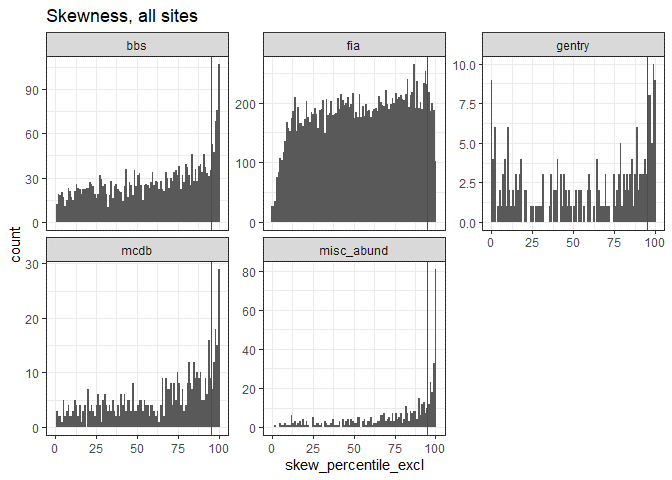
##### 

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



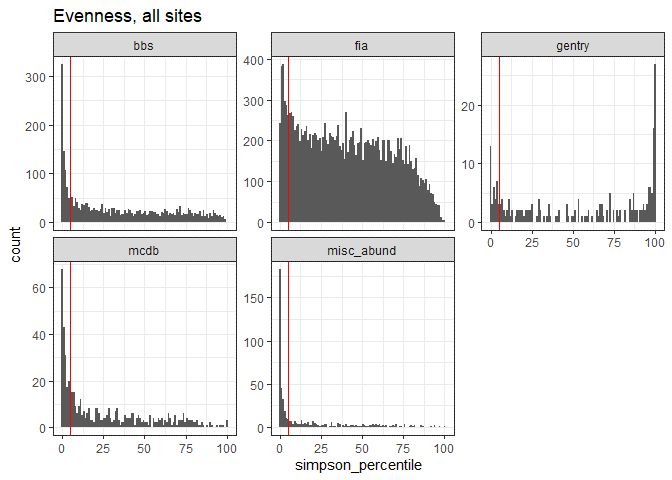
For every sample drawn from the feasible set (left column), we calculate the skewness (color scale) or evenness (not shown). These values generate the distribution of expected statistic values from the feasible set (right column). The ratio of the range encompassed in the 95% interval (space between 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



Dotted line is 95.

Excludes communities for which S < 3



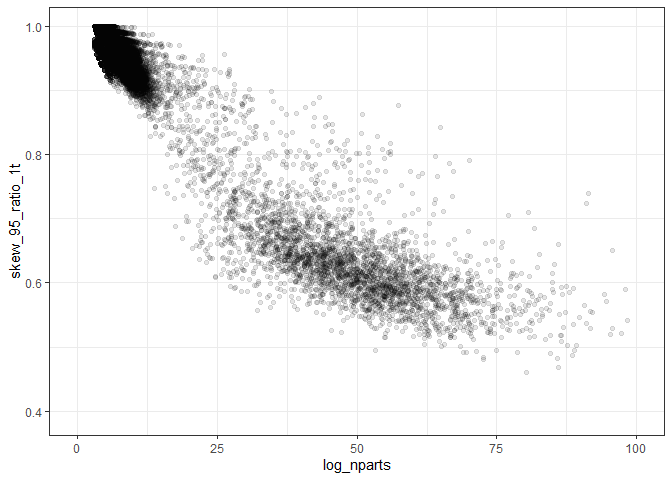
Dotted line is 95.

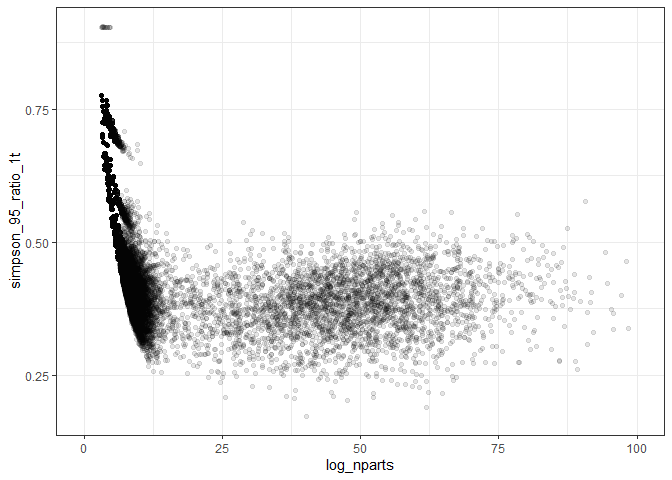
##### Table 1: Percentile results

| **dat** | **proportion\_skew\_high** | **nsites\_skew** | **proportion\_even\_low** | **nsites\_even** |
| --- | --- | --- | --- | --- |
| bbs | 0.1301839 | 2773 | 0.2596466 | 2773 |
| fia | 0.0542077 | 18300 | 0.0939657 | 18113 |
| gentry | 0.1883408 | 223 | 0.1517857 | 224 |
| mcdb | 0.1582868 | 537 | 0.3542435 | 542 |
| misc\_abund | 0.3455285 | 492 | 0.5959184 | 490 |

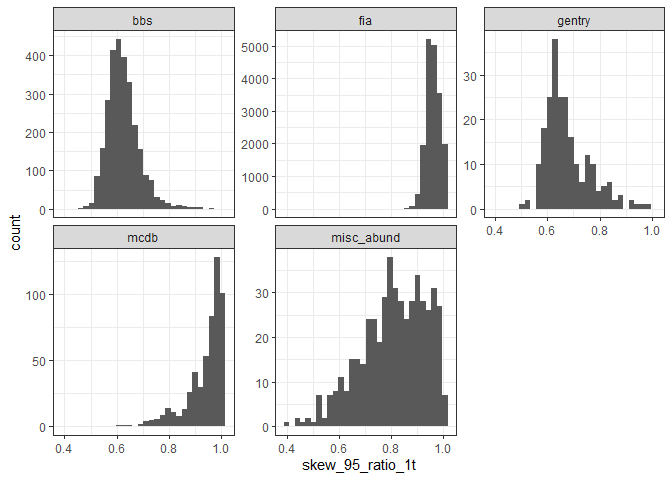
| **fia\_yn** | **proportion\_skew\_high** | **nsites\_skew** | **proportion\_even\_low** | **nsites\_even** |
| --- | --- | --- | --- | --- |
| fia | 0.0542077 | 18300 | 0.0939657 | 18113 |
| not fia | 0.1634783 | 4025 | 0.3072723 | 4029 |

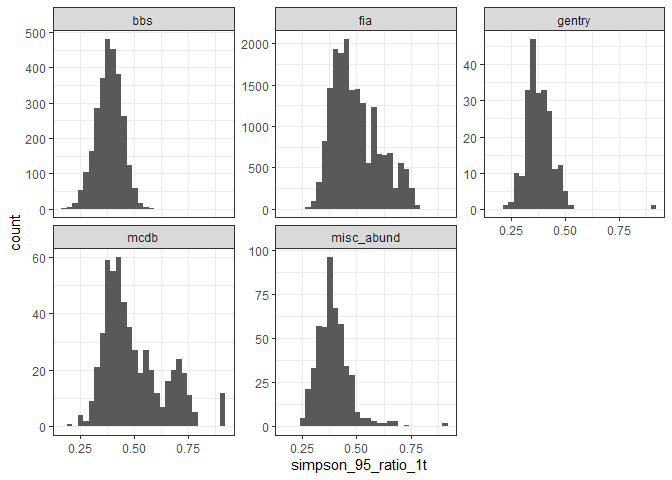
##### Figure 5: 95% intervals vs. size of feasible set



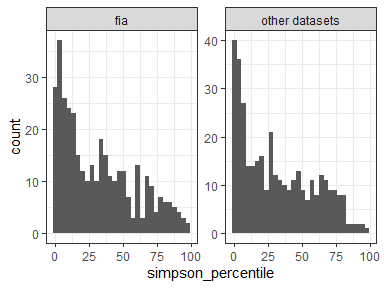
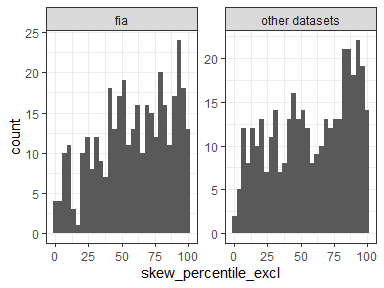


##### Figure 7: Distribution of 95% intervals by dataset





##### Figure 8: Direct comparison of FIA and similarly sized sites



##### Table 2: Percentile results comparing FIA to direct counterparts

| **fia\_yn** | **prop\_skew\_high** | **prop\_even\_low** | **nsites\_skew** | **nsites\_even** |
| --- | --- | --- | --- | --- |
| fia | 0.0808625 | 0.1780822 | 371 | 365 |
| other datasets | 0.0862534 | 0.2109589 | 371 | 365 |