**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 this remarkably general pattern in species abundances to test numerous theories designed to determine which biological processes are most important for structuring assemblages of species (McGill et al 2007). However, using the species abundance distribution to distinguish amongst theories has proven frustrating because many theories produce similar shapes (McGill et al 2007) and even experimental manipulations generate little variation in the shape of this distribution (Supp and Ernest 2014). After decades of use as a theoretical benchmark, the usefulness of the SAD for assessing the processes structuring ecological communities is currently unclear.

Increasing evidence for the role of statistical constraints in generating the SAD (refs) further complicates our understanding of the SAD as a *biological* pattern. Accumulating evidence suggests that statistical constraints can account for the most striking feature of the species abundance distribution – the hollow curve. Fundamentally, the SAD describes 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 it be uniformly distributed (i.e. equal numbers of individuals per species), the hollow curve we see in nature is surprising. However, uniformity is not necessarily the correct baseline. Just like 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. Maximum Entropy Theory of Ecology (METE); Harte et al 2008, Harte 2011) or combinatorics (i.e. ‘the feasible set’; Locey and White 2013), randomly partitioning individuals among species can, on its own, generate realistic hollow curves when constrained by a community’s values of S and N. Because such statistically generated SADs are excellent empirical fits to those seen in nature (Harte et al 2011, White et al 2012), it is reasonable to expect statistical constraints alone may account for the widespread nature of the hollow SAD curve.

If SADs are statistically inclined to be hollow curves even without invoking biological process, it is no surprise that we have struggled to interpret the hollow curve in biological terms. However, just because the most readily distinctive feature of the SAD may be statistically determined, it does not necessarily follow that we cannot use the SAD to understand biology. 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 evaluating SADs not in terms of their absolute shape, but in terms of their shape relative to what we would expect simply due 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 developing metrics and computational approaches that allow us to quantify and interpret, across many different types of communities, any deviations that may exist.

Our ability to determine whether an empirical SAD is unlikely when compared to its statistical baseline may be strongly impacted by the size (in terms of S and N) of the community. This is because the statistical baseline is in fact a *distribution* of outcomes from random processes, of which some outcomes are more likely than others to emerge at random. In the study of complex systems, predictions for the most-likely outcomes are most easily and confidently derived “in the limit” of aggregating over large numbers of components – for example, atoms of gas in a vessel (Frank 2009, Harte et al 2011). As the number of components in the system becomes very large, the distribution of possible outcomes for the system clusters tightly around a single highly-likely result, making even that (Jaynes 1957). Unlike vessels of gas, ecological communities are much more finite in size. Smaller systems may have a relatively broad distribution of possible outcomes, and even observations that are quite different from the *most* likely outcome will not necessarily appear highly unlikely. When this occurs, we cannot be sure if an observation that differs from the most-likely outcome, but falls within the realm of statistical possibility, is a result of random sampling or a reflection of nonrandom process. The questions of whether SADs for ecological communities are susceptible to these small-size effects, and what community sizes (i.e., values of S and N) are most affected, remain open.

Here we set out to establish if, and in what ways, empirical SADs deviate from statistical baselines derived from random partitions of individuals among species. Additionally, we explore how community size affects the breadth or narrowness of the distribution of randomly-generated outcomes, and whether this appears to modulate our ability to detect deviations between empirical observations and randomness. We build upon the combinatorics approach developed by Locey and White (2013) to obtain the randomly-expected distributions of outcomes for the SAD based on the number of species and number of individuals. Because exhaustively exploring the range of possible outcomes for realistic values of S and N is computationally prohibitive, we developed a new algorithm to draw samples at uniform from this outcome space and characterize distributions of randomly expected outcomes across broad ranges of S and N. For [# communities] communities spanning X Y and Z taxa, we compare *observed* SADs to their corresponding random expectations and evaluate 1) if the shapes of observed SADs are consistently unusual relative to these random outcomes, 2) how the narrowness of the distribution of random outcomes varies over gradients in *S* and *N*, and 3) whether variation in the narrowness of this distribution is associated with the apparent deviation, or lack thereof, between observation and randomness.

**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 on 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 further filtered these data to remove exceptionally large or trivially small communities because characterizing the statistical constraint on the SAD is computationally intractable for the very largest communities, and is trivially uninformative for the very smallest ones. We removed 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 t, whichcomprises roughly 100,000 communities, 90,000 of which speciesAFIA were included in the analysisFinally, for any site 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 Abundance Database and Miscellaneous Abundance Database collected over longer timescales that cannot be disaggregated, with an average temporal scale of X. After filtering multiple years of data and exceptionally large or small communities, our final dataset consisted of

*the statistical baseline*

We use the concept of the *feasible set* to characterize the statistical baseline for the SAD (Locey and White, 2013). For a given number of individuals *N*,there is a finite possible set of unique ways to partition those individuals into *S* species. The feasible set is the complete set of these unique partitions (or elements), each of which is considered an equally likely result of randomly dividing *N* individuals into *S* species. 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 considered distinguishable from each other (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 elements of the feasible set, but *(1, 3, 5)* and *(3, 1, 5)* count as only one element of the feasible set. 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 approximate reality (Locey and White 2013).

While it is possible to exhaustively list the elements of the feasible set for very small values of S and N, it is necessary to use sampling to explore the feasible set for most ecologically-relevant values of S and N. We developed an algorithm to efficiently and uniformly sample feasible sets even for large values of S and N. We implemented this algorithm in an R package, available on GitHub at [www.github.com/diazrenata/feasiblesads](http://www.github.com/diazrenata/feasiblesads); details of the sampling methodology are available at <https://github.com/ha0ye/feasiblesads/blob/algo-vignette/vignettes/sampling_algorithm.Rmd>. For every community in our database, we uniformly drew 5000 samples from the feasible set to characterize the distribution of statistically probable shapes for the SAD. We filtered the 5000 samples to unique elements. Note that for small values of S and N – for which there may even be fewer than 5000 unique elements in the feasible set – it can be impossible or highly improbable to randomly draw 5000 *unique* samples from the feasible set. For large communities, all 5000 are usually unique.

Each element of the feasible set is a SAD shape, and the feasible set is thus a statistical distribution of SAD shapes that provide a baseline for understanding what SAD shapes are most likely to emerge from random processes constrained only by S and N. To assess whether observed SADs deviate from this baseline, we compared the shapes of empirically-observed SADs to the distributions of shapes present in their feasible sets. To quantify the shapes of both observed SADs and SADs from the feasible sets,

we calculated the skewness and Simpson’s evenness for each SAD. Metrics related to the evenness of the distribution of abundances across species 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 chose these two specific metrics because Simpson’s evenness is widely used in ecology and skewness is a core moment of distributions. Calculating Simpson’s eveness and skewness for each sample from the feasible set generates a distribution of values for each metric that reflected the likely evenness or skewness values for a random SAD with the S and N used to generate the feasible set. Note that skewness [breaks under specific circumstances – s < 3, all abundances equal], and we exclude those cases from analyses of skewness.

The feasible set’s distribution of Simpson’s evenness or skewness was used to assess whether observed SADs were statistically unlikely given their values of S and N. We assessed whether the observed Simpson’s evenness or skewness deviated from its feasible set by calculating the percent of values in the sample distribution less than or equal to the observed value. This percentile rank is comparable across different community sizes, allowing broad-scale assessment across wide ranges of S and N. If observed SADs are just random draws from their feasible sets, we would expect 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.

The distributions of summary statistic values from samples from the feasible set also allowed us to describe the relative narrowness or breadth of the statistical constraint over gradients of S and N. We quantified the narrowness of a feasible set distribution as the ratio of the range of values encompassed within a one-sided 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 vary considerably in their actual values, means, ranges, and standard deviations. Results from this metric correspond qualitatively to more computationally-intensive approaches to measuring the self-similarity of the elements of feasible sets (Supplement).

*Effects of rarefaction*

Generally, it is logistically impossible to conduct exhaustive surveys of ecological communities, and even the most intensive surveys of nature are likely to miss some species – particularly rare or cryptic ones. To test whether missing rare species influenced the relationship between observed communities and their feasible sets, we ran our analyses on both raw data and on SADs adjusted via rarefaction. For each raw SAD, we took the mean number of estimated species from [] richness estimators, and added species to reach this mean estimated richness. We reasoned that species missed during sampling are likely to be rare, and added a single individual each for each added species. This also allowed us to explore the consequences of rarefaction while making the smallest possible changes to S and N.

**Results and Discussion**

We found evidence that the shape of the SAD for a range of real ecological communities are statistically unlikely when compared to their feasible set. 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:_Skewness_1), [Figure](#_Figure_4:_Evenness)). 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, w, where Why the FIA dataset shows a different pattern is unclear, but it may be due to their small size, which we explore further below. For the four other datasets, however, our results suggest that …

of something pushing observed SADs into unusual states of of our variation in this signal at the individual SAD level. Some of this variation may be due to differences in the strength of ecological processes structuring communities, as might be expected from theory, but there are also statistical issues that could also underlies this heterogeneity. Adjusting for rarefaction increased the strength of this signal: for these datasets, 18% of adjusted SADs are more skewed than 95% of their feasible sets, and 38% are less even. In nearly every case, the percentile scores for rarefaction-adjusted SADs were more extreme than for the raw SADs ([Figure](#_Figure_5:_Rarefaction)). If anything, incomplete sampling seems to cause us to *under* estimate the deviation between observation and randomness.

in how communities compare to their feasible set due to statistical issues with community size. Community size – in terms of *S* and *N* – may affect our ability to distinguish between deviations and randomness via its effect on the variability of forms represented in the feasible set. Additionally, FIA communities are among the smallest in our database, and we saw little evidence of these communities deviating from their feasible set. This would be consistent with the concern than small communities may not have enough possible states, or elements in their feasible set, to generate highly resolved distributions with strong definitions of the mostly likely state of the system.

When we examined the shape of the feasible set distributions, for small versus large communities, we found that small community feasible sets have broader distributions of evenness, and especially skewness, than those for large communities ([Figure](#_Figure_8:_95%)). For communities of the sizes represented in the FIA dataset, the 95% interval of skewness values often encompasses more than 80% of the entire range of values; for larger communities, the 95% interval spans closer to 60% of the full range. This is consistent with concepts from statistical mechanics: large communities have many components that can be arranged in many ways, and most of these arrangements cluster around a relatively specific highly-likely state. If an observation differs even a small amount from this most-likely state, it is readily detectable as highly unlikely to have occurred by chance. Small communities have relatively few possible arrangements and relatively broad distributions of likely shapes. Observations may deviate from the *most*-likely form, but only the most extreme deviations will be highly unlikely given the breadth of the corresponding probability distribution. Although we cannot definitively show that small community size accounts for the pronounced difference in deviations between FIA and other datasets, we think it the most likely explanation, given that we observe broad statistical constraints for small communities and that we see similarly less-common deviations in small communities in the non-FIA datasets.

If this is indeed the case, it means that small-community considerations are relevant for ranges of S and N that are quite common in ecology. The FIA communities range in size from x to y species and x to y individuals. These are by no means hard thresholds, but they 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 in 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 statistical constraint is relatively broad, and may even help us develop a type of power analysis to find the smallest-detectable deviation for a given community size.

In general, our results suggest that the shape of the SAD is not entirely a statistical artefact – that ecological SADs exhibit a particular, highly uneven, form that may reflect the signal of ecological processes that structure communities. Figuring out what processes are generating these deviations in ecological communities will require further exploration. One obvious next step is assessing whether existing theories (e.g. neutral theory, METE, ???) and common functional approximations (e.g., logseries, exponential, and log normal) can predict deviations from the feasible set. Our results also suggest that whatever processes are structuring communities, they are more likely to be ones that push abundance distributions towards a more uneven state rather than processes that end up spreading individuals more evenly across species.

While our results provide support for ecological processes structuring species abundance distributions, they also suggest there may be limits to our ability to distinguish between randomly structured communities and ecological structured ones – particularly in the case of small communities. If this is indeed the case, it means that small-community considerations are relevant for ranges of S and N that are quite common in ecology. In our study, we commonly saw broad feasible set distributions was in the range of x to y species and x to y individuals. These are by no means hard thresholds (and were driven primarily by FIA), but 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 in 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 statistical constraint is relatively broad, and may even help us develop a type of power analysis to find the smallest-detectable deviation for a given community size.

It is also important to recognize that there are multiple plausible approaches to defining the statistical baseline for the SAD, of which we have taken only one. 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 the use of one baseline over another, but comparing the performance of 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 constraints 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





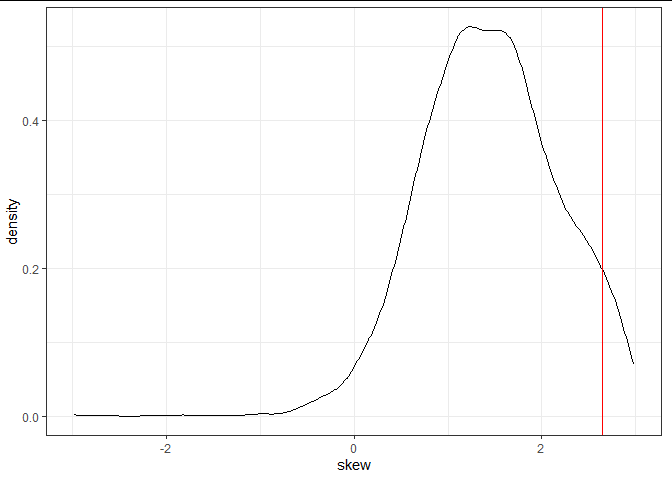


##### 

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

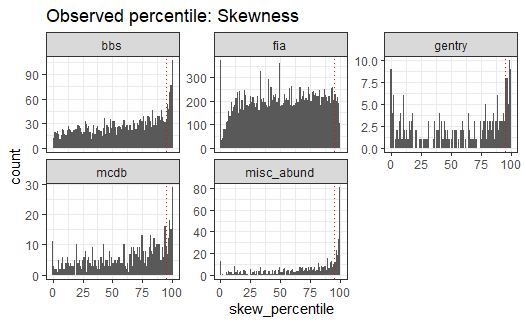
Range of values for 95% density

÷



Full range of values

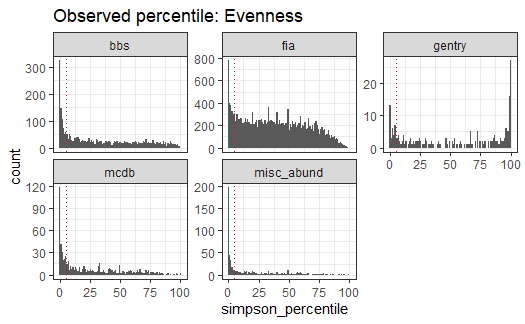
##### Figure 3: Skewness percentile results



Dotted line is 95.

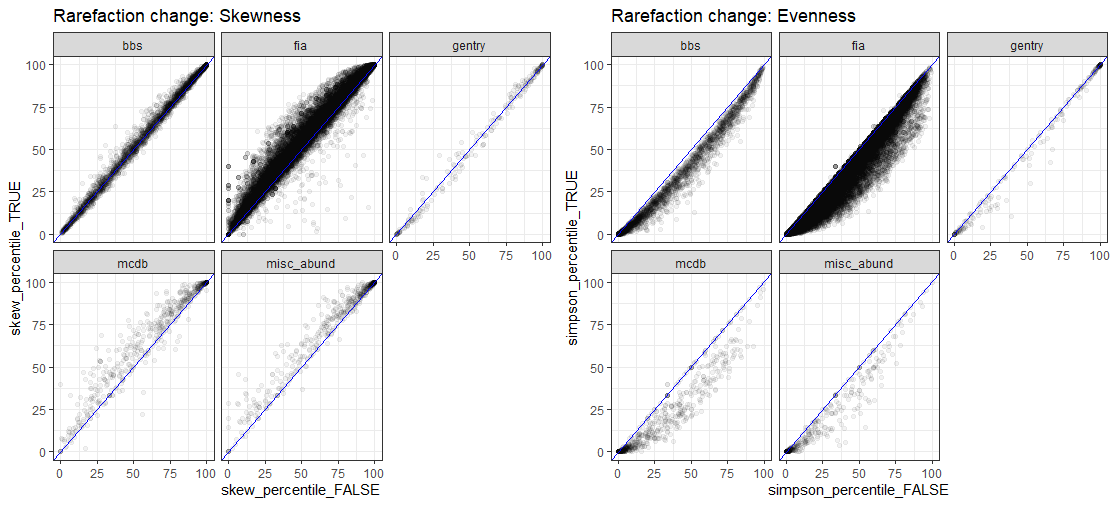
Excludes communities for which S < 3

##### Figure 4: Evenness percentile results



Dotted line is 95.

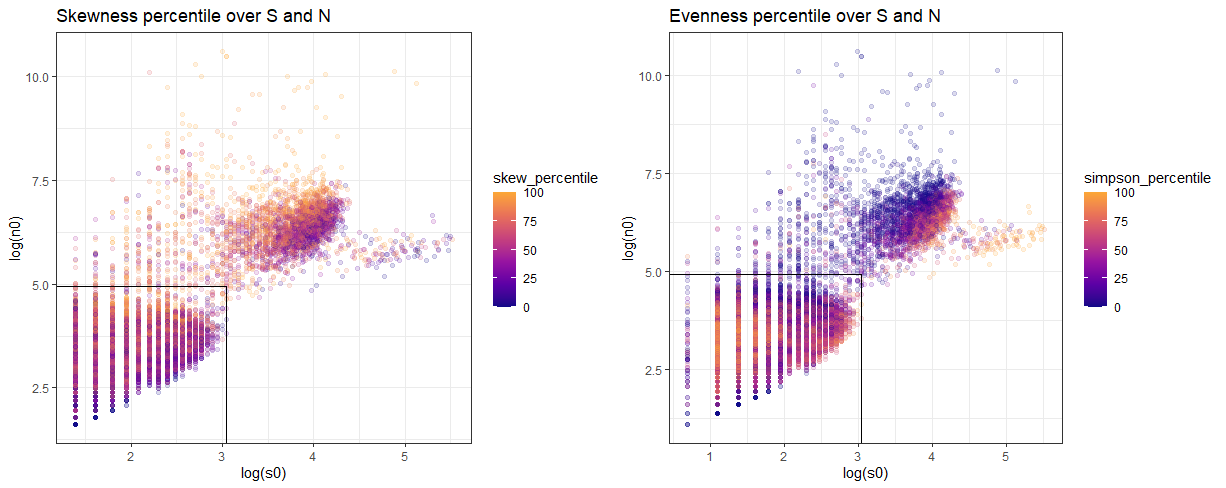
##### Figure 5: Rarefaction effects



Blue line is 1:1

Skewness excludes s < 3

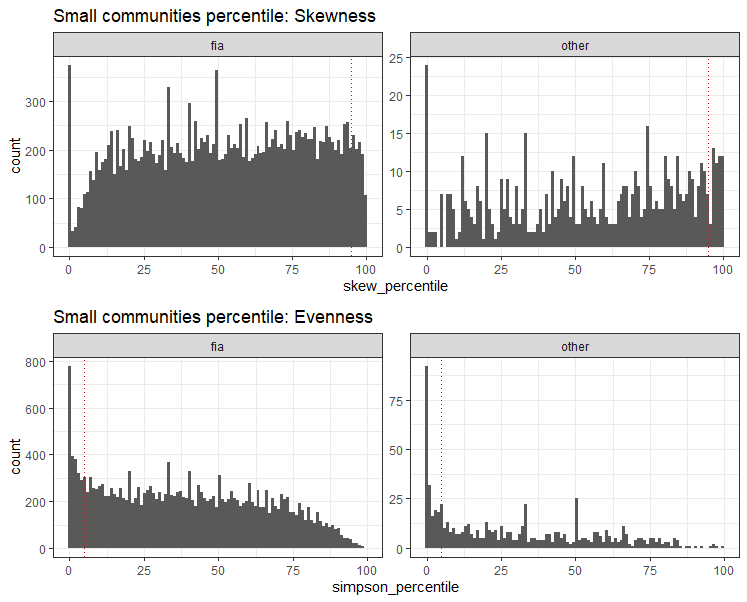
##### Figure 6: Percentile scores over S and N



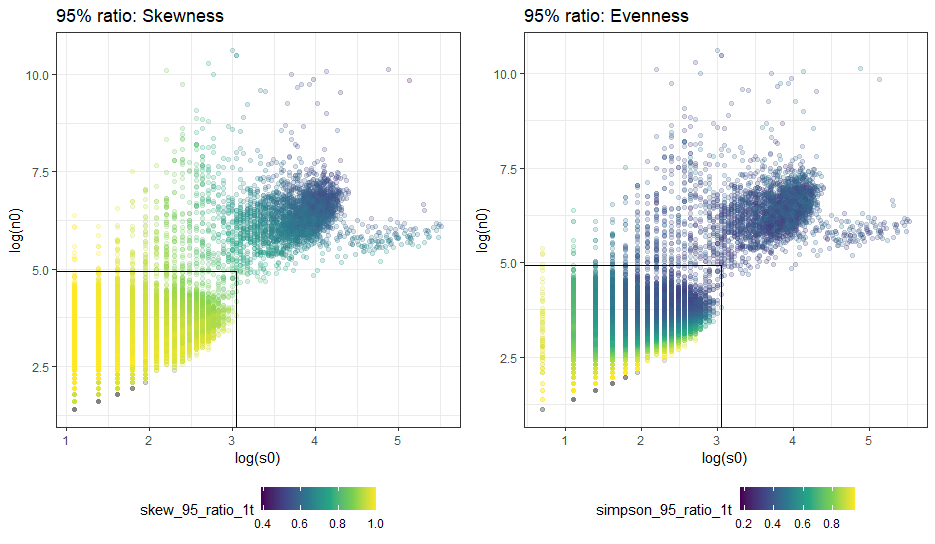
Skewness excludes s < 3

Black line outlines FIA maximum s0 and n0

##### Figure 7: Percentile outcomes for small communities



##### Figure 8: 95% ratio over S and N



Skewness excludes s < 3

Black line is FIA maxima