**Introduction**

The species abundance distribution (SAD) is one of the very few ecological patterns whose consistency in shape makes it akin to 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). Tempted by this remarkably general pattern, community ecologists have used the species abundance distribution 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 infer underlying processes has proven frustrating because there is so little variation in the shape of this distribution, even in response to experimental manipulations (Supp and Ernest 2014), and many theories produce similar predictions that perform comparably well when compared to observed distributions (McGill et al 2007).

Ecologists have generally focused on explaining the species abundance distribution as emerging from ecological processes, but accumulating evidence suggests that statistical constraints may be involved in generating the most striking feature of the species abundance distribution – the hollow curve. Mathematically speaking, we can think of an SAD as an unordered vector of S values that sum to N, where S and N are the total number of species and individuals in a community. If our implicit null expectation for such a distribution is that it be uniformly distribute, the hollow curve we see in nature is indeed surprising and begs for explanation. However, multiple efforts to characterize the most-likely random expectation for the SAD have converged to find that in general, an appropriate statistical baseline for the SAD is in fact *a hollow curve*. Analogous to the way a Gaussian distribution emerges in the limit of many samples around a population mean, power-law or logseries distributions emerge as statistical attractors for abundance distributions (Frank 2009, 2019). The Maximum Entropy Theory of Ecology (METE) uses methods derived from statistical mechanics to find the most-likely form for numerous ecological distributions, given sparse information about broad community properties and minimal assumptions regarding ecological mechanism (Harte et al 2008, Harte 2011). Even parameterized only with S and N, METE predicts a log-series SAD that is an excellent fit to empirical SADs (Harte et al 2011, White et al 2012). Finally, we can use combinatorics to explore the set of *possible* forms, or feasible set, for an SAD with a particular S and N. Hollow curves, similar to the ones seen in empirical SADs, dominate most feasible sets for biologically-relevant combinations of S and N (Locey and White 2013). Given minimal assumptions about biological process, and instead using S, N, and our understanding of the most-likely form of a vector with S elements that sum to N, we therefore have good reason to expect most SADs to be hollow curves.

If SADs are statistically inclined to be hollow curves before biology even enters the picture, it is no surprise that we have struggled to interpret the hollow curve in biological terms. It may be more informative to evaluate SADs not in terms of their absolute shape, but in terms of their shape *relative to what we would expect simply due to S, N, and the statistical constraint* (Frank 2009, Locey and White 2013). While the general hollow-curve shape may be a statistical artefact, it is still possible that biological processes push real SADs away from their most likely outcomes. This could result in relatively subtle, but meaningful, deviations between observed SADs and their statistical expectations (Locey and White 2013, Harte and Newman 2014). If we detect such deviations, we can evaluate theories in terms of how well they predict the *deviations,* which is to say, how much additional predictive power they give us beyond what is inherent to S and N (Harte and Newman 2014, Xiao et al 2016). If we cannot distinguish between observed species abundance distributions and their statistically-most-likely forms, there may be little information we can extract from the SAD alone – either because it does not contain appreciable signal of biological process, or because the range of highly-likely forms for the distribution obscures that signal.

Our capacity to detect deviations between observations and randomness may depend strongly on the distribution of statistically-likely outcomes, which in turn depends on the size of the system in terms of S and N. Generally, in the study of complex systems, predictions for the most-likely state of a system 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 particles in the system becomes very large, the set of *possible* states of the system clusters tightly around a single highly-likely state, meaning any other outcome would be a true statistical surprise (Jaynes 1957). Unlike vessels of gas, ecological communities do notalways have effectively infinite subcomponents, and may have not have such narrowly-defined random expectations. This can interfere with our ability to detect deviations from randomness. If there is a narrowly-defined, overwhelmingly most-likely form for a particular SAD, even a relatively small departure from this outcome is highly unlikely to occur at random. However, if the probability distribution of possible forms for an SAD is relatively broad, even observations that are quite different from the *most* likely outcome are not necessarily 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. We expect this situation to be more common in smaller communities, but *a priori* we do not know the specific ranges of S and N for which it will be relevant.

We set out to establish if, and in what ways, we can detect consistent deviations between empirical SADs and their statistical constraints. If present, consistently detectable deviations would provide new leverage for interpreting SADs in biological terms. Such deviations likely betray non-random processes at work, and we may be able to use the deviations to evaluate theoretical predictions and make comparisons between empirical systems. The *absence* of detectable deviations would not necessarily be evidence of an overall lack of structural processes, but would mean that it is likely to be quite difficult to extract biological information from the SAD. Especially if such a lack of detectable deviations corresponds with small communities that have relatively vaguely-constrained SADs, we suspect it reflects the limits of our ability to make confident statistical distinctions in those contexts.

For [# communities] communities spanning X Y and Z taxa, we characterize the distribution of statistically-likely forms for the SAD based on the number of species and number of individuals. We then compare *observed* SADs to their corresponding statistical baselines and evaluate 1) whether observed SADs are consistently unusual given their statistical expectations, and 2) whether our capacity to detect deviations seems to vary systematically over *S, N,* and corresponding variation in the narrowness of the statistical constraint.

**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 feasible set (White et al 2012, Baldridge et al 2014). It includes data from trees in the Forest Inventory and Analysis and Gentry plots, the North American Breeding Bird Survey, the Mammal Community Abundance Database, and a variety of less commonly sampled taxa in the Miscellaneous Abundance Database.

For datasets with observations from multiple years, we follow White et al (2012) and Baldridge et al (2014) and analyze only a single year of data. The Mammal Community Abundance Database and Miscellaneous Abundance Database include data collected over longer timescales that cannot be disaggregated, with an average temporal scale of X.

Our approach to characterizing the statistical constraint on the SAD becomes computationally intractable for the very largest communities, and becomes trivially uninformative for the very smallest communities. 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.

The FIA database contains roughly 100,000 communities, of which 90,000 have fewer than 10 species and X individuals. Rather than analyze all of these small communities, we randomly selected 10,000 small communities to include in the analysis. We analyzed all communities in FIA with more than 10 species. In all, we analyzed X communities encompassing X taxa, with S and N ranging from 2 to X and X to X, respectively.

*the statistical constraint*

Following Locey and White (2013), we use combinatorics to characterize the statistical constraint on the SAD. This allows us to describe both the most-likely form for the SAD, and the *distribution* of possible forms, without additional mathematical machinery, assumptions, or parameterization beyond S and N. For any given number of individuals *N*,there is a finite set of unique ways to partition those individuals into *S* species, or the “feasible set”. We define the SAD as unordered, and count ties in abundance as interchangeable; we note that making other assumptions regarding the distinguishability of species would change the composition of the feasible set. A uniform sample of elements from the feasible set reflects the distribution of possible forms for an SAD, and – to the extent to which the elements of the feasible set converge towards one overall shape – illustrates both the shape of the most-likely form and the narrowness of the constraint. Uvalues of

*Effects of S and N on the size and narrowness of the feasible set*

For every community in our database, we calculated the number of elements in the feasible set for an SAD with the corresponding S and N. We drew 5000 draws from every feasible set, and filtered this 5000 to unique samples. For large feasible sets, all 5000 are usually unique; for small ones, it becomes or impossible to find 5000 unique samples.

To illustrate how the narrowness of the statistical constraint varies with S, N, and the size of the feasible set, we examine the distribution of shapes represented in our samples from the feasible set. We summarized the shapes present by calculating the skewness and Simpson’s evenness for every sample from the feasible set, and examined the distribution of values for these summary statistics. The actual values achieved for both statistics depends on S and N, so we focus on the shape of the distribution of values instead of on the values themselves.

Specifically, we compared the rangeof values present in the 95% interval of values to the full range across all samples. A narrow statistical constraint translates into a steep, narrow peak in the distribution of summary statistic values. Most of the density is concentrated around that peak, and the 95% interval encompasses a relatively small proportion of the overall range of values. In contrast, a relatively vague statistical constraint translates to a broad, gentle peak in the distribution of summary statistics. The 95% interval includes a larger proportion of the range in values present in the distribution. We calculated the ratio of the 95% range to the total range, and use this ratio to describe the narrowness of the statistical constraint for every combination of S and N sampled. This approach is certainly imperfect, but its findings agree with a separate self-similarity analysis not presented here (see supplement).

*Comparing observed SADs to their feasible sets*

We compare the forms of observed SADs to their feasible sets by comparing summary statistic values for the observed SADs to the distribution of values from samples from the feasible set. We again used skewness and Simpson’s evenness as our summary statistics. The range of values we observe for both statistics depends on S and N, so in order to draw conclusions over communities with different S and N we must focus on the percentile value, rather than on the raw values of the statistics. We calculated percentile rank as the percent of values in the distribution less than or equal to the observed value. Note that the precise meaning of this value differs subtly depending on whether the distribution of samples from the feasible set represents a true distribution of *samples* or an exhaustive account of all elements in the feasible set. If the samples are exhaustive, the percent of samples less than or equal to the observed value may be 100, but can never be 0 – even if the observed value is the lowest value possible, it is still less than or equal to itself. However, if sampling is not exhaustive, we may obtain a percentile value of 0 if the observed value is less than the lowest value we came across via sampling.

At random, we would expect the percentile values to be uniformly distributed from 0 to 100. If observed SADs are consistently more skewed or even than their feasible sets, the percentile values will be disproportionately concentrated towards the extremes. If, as we suspect, smaller communities have relatively vaguely-defined statistical constraints, we may not perceive consistent deviations from randomness in those communities. We present results for percentile values aggregated over all datasets, and sub-set according to the size of the feasible set and the narrowness of the statistical constraint.

*Effects of rarefaction*

Whenever we survey the species present in a real system, we are likely miss some species – particular rare or cryptic ones. To test whether this affects our results, 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 allows us to explore the consequences of rarefaction while making the smallest possible changes to S and N.

**Results**

*Effects of S and N on the narrowness of the statistical constraint*

Over the range of S and N in our data, both the number of elements in the feasible set and the narrowness of the constraint increase with S, N, and the ratio of N:S ([Fig 1](#_Figure_1:_S0,), [Fig 2](#_Figure_2:_Self-similarity)). This is consistent with our expectations drawn from statistical mechanics (Jaynes 1957, Harte et al 2011). For large communities with large feasible sets, we have a very specific and narrowly-defined statistical constraint. We can be confident that even slight deviations from this baseline are very unlikely outcomes. However, we may not be able to make such distinctions for small communities with smaller and less narrowly-defined feasible sets. For these feasible sets, even an outcome that is quite dissimilar to the *most* likely outcome is still within the realms of statistical possibility. Importantly, we see relatively broad statistical constraints even for values of S and N that are common in ecology. To the extent that this vagueness limits our ability to detect deviations from the random baseline, we may stand to learn from SADs only for relatively large ecological communities.

*Observed SADs relative to their constraints*

Observed SADs are often very skewed and uneven compared to their feasible sets, but the strength and even existence of this phenomenon vary across the datasets we analyzed (Fig 3, Fig 4). Excluding FIA, 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.

In contrast, the FIA percentile scores are near uniformly-distributed for skewness (5% of observations are highly skewed), and much noisier than any of the other datasets for evenness (11.5% of observations are highly uneven). We suspect that this is primarily an effect of small community size: the FIA communities are the smallest in our database. There is not a strong difference between FIA communities and other small communities in the distribution of percentile values (Fig 5), although note that the vast majority of communities in this size range are FIA. Communities in this size range have extremely broad statistical constraints (Fig 6), especially for skewness, which may make it impossible to confidently detect deviations.

The Gentry plots are also unusual in that they are often *more* *even* than their feasible sets. This appears to be restricted to plots with a very low ratio of N/S (<1.5), for which the entire feasible set is forced to be extremely even.

*Effect of rarefaction*

SADs corrected for rarefaction are almost always *more* unusual, if anything, compared to their feasible sets than the raw distributions. In nearly all communities, the rarefied percentile value for evenness is lower than for the raw distribution, and the percentile value for skewness is higher than the raw one (Fig 7). Correcting for rarefaction, overall 18% of non-FIA SADs are more skewed than 95% of their feasible sets and 38% are less even than 95% of their feasible sets, while 9% and 17% of FIA sites are highly skewed and uneven, respectively. Correcting for rarefaction may increase the apparent differences between observed SADs and their feasible sets partially because we added new species as singleton species; adding singletons to a distribution without changing the rest of its shape necessarily increases skewness and decreases evenness. That said, we *should* expect failure to observe rare species to artificially truncate the rare tail of the SAD.

**Discussion**

Empirically observed SADs often deviate pronouncedly from their statistical constraints, but this is by no means a rule even for the largest communities. We do not find such strong deviations for small communities, including the Forest Inventory and Analysis plots, but we are hesitant to interpret this as a biologically-driven phenomenon. We think it more likely that the characteristics of the statistical constrainton these small communities limit our capacity to detect any but the most extreme deviations.

large communitiesmore– but note that it is hardly uncommon for the observed SAD to fall well within our random expectation. This range of variation in *deviation* between observation and random expectation may provide new leverage for evaluating theory and drawing contrasts between communities. If a theory can correctly predict whether a community will or will not deviate, or better still the magnitude of the deviation, it is telling us considerably more than we would be able to infer from S and N alone. Similarly, we may be able to identify differences between communities that determine whether their SADs are detectably different from the constraint.

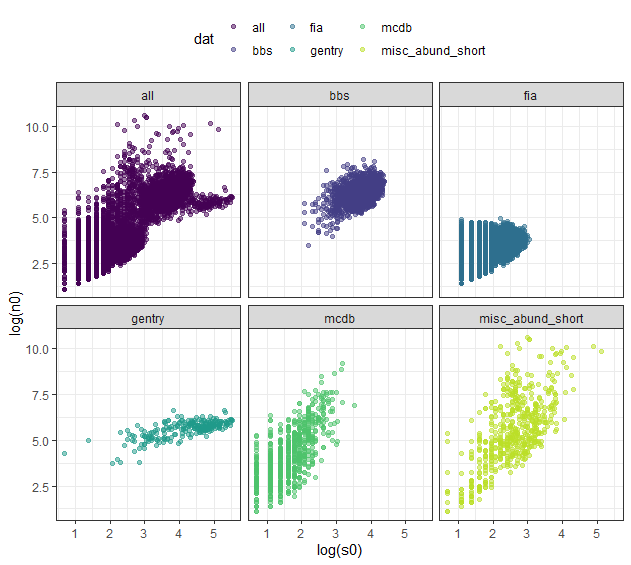
The picture is less straightforward for relatively small communities, including the Forest Inventory and Analysis dataset. We detect no difference between observation and the random expectation of the constraint in skewness, and a much less pronounced – but still detectable – trend when it comes to evenness. It is technically possible that these small communities differ in some substantive way from large communities, and that large ones are structured by strong nonrandom processes while small ones are more nearly random. However, we think it more likely that our failure to detect deviations for small communities is simply a reflection of the underlying characteristics of the statistical constraints on these communities. Generally, small systems are problematic for approaches inspired by statistical mechanics, as ours is here. The statistically most-likely state for these systems is relatively broadly defined, and this can make even nonrandom outcomes difficult to distinguish from outcomes that could reasonably emerge at random. Consistent with this explanation, we find that the distributions of summary statistic values from small feasible sets are extremely broad. Unless we can develop more sensitive methods for detecting deviations, or a lack thereof, for these small communities, we should expect to learn the most by focusing on SADs from relatively large communities.

The SAD is a case study that illustrates how important it is to consider statistical baselines when we interpret “patterns” in ecology and especially macroecology. Like all ecological distributions, the SAD emerges from a suite of biological *and statistical* processes. In this case, the statistical constraint on the SAD thwarts our intuition for what constitutes a surprising form for the distribution. Accounting for the statistical constraint reveals more subtle patterns of variation, which may yield much more insight into biology than we have so far managed to extract from the famous hollow curve.

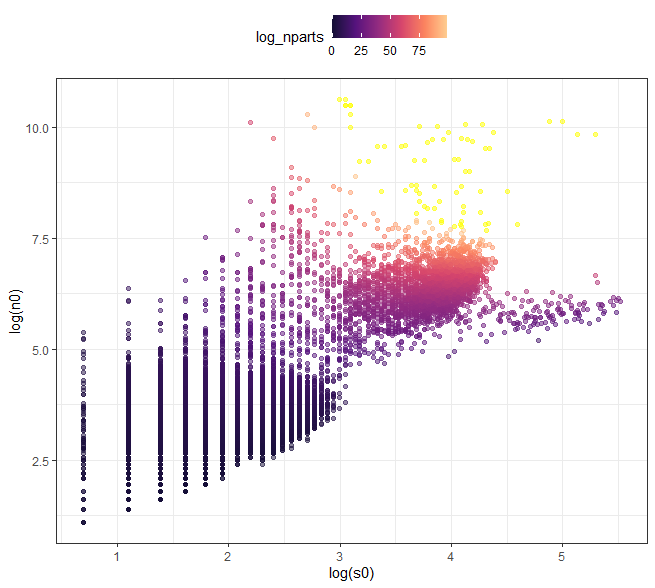
It is relatively intuitive to demonstratethis phenomenon for the SAD, but the general principle may be equally important for other, less tractable, distributions. Because the SAD involves partitioning discrete values into distinct categories, we can characterize the constraint fairly simply via combinatorics. Defining and accounting for the statistical constraints on more complex, often continuous, distributions poses additional challenges and may necessitate borrowing tools from other disciplines. If the SAD is any indication, rising to those challenges and incorporating the statistical baseline into our analyses of other ecological “patterns” is both necessary and richly informative.

**Figures**

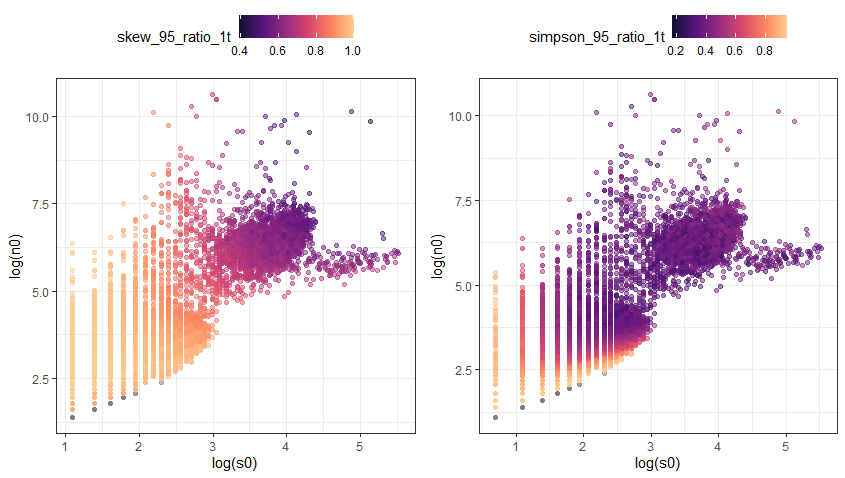
Figure 0: Distribution of datasets in S and N space



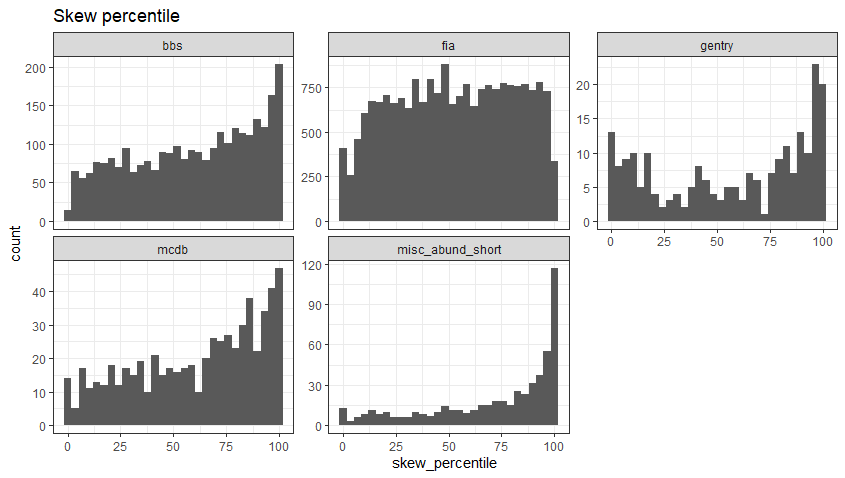
##### Figure 1: Number of elements in the feasible set



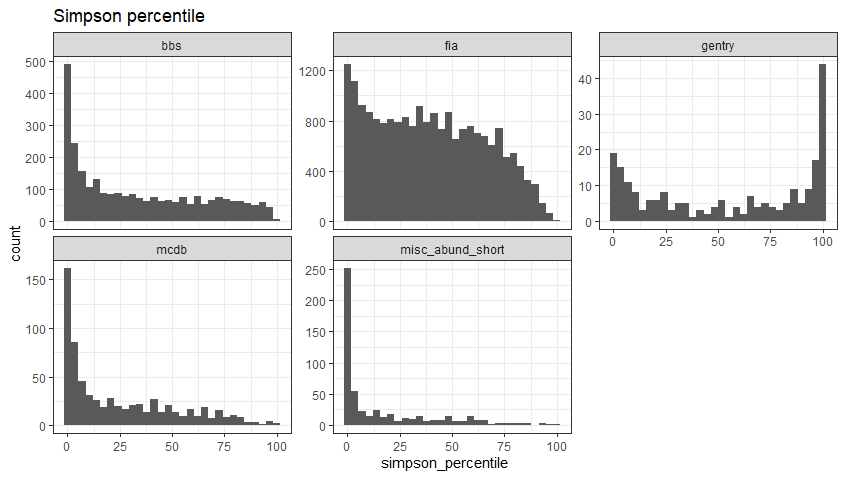
##### Figure 2: Narrowness of constraint over S and N



##### Figure 3: Skewness overall



##### Figure 4: Simpson overall









##### Figure 5: Within the FIA size range, skewness and evenness

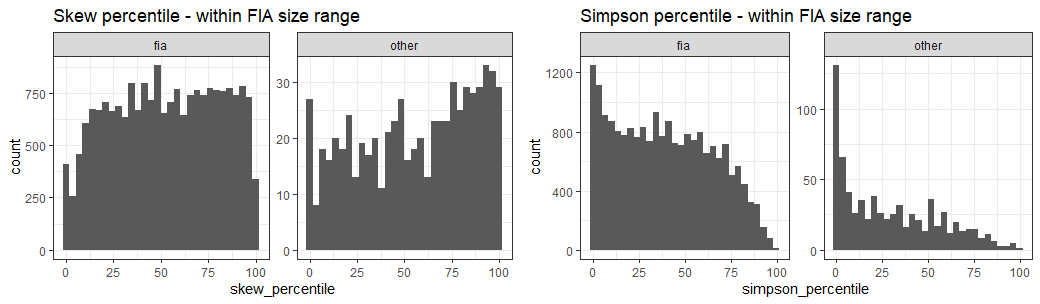


Figure 6: Narrowness of statistical constraint for FIA sized and larger data

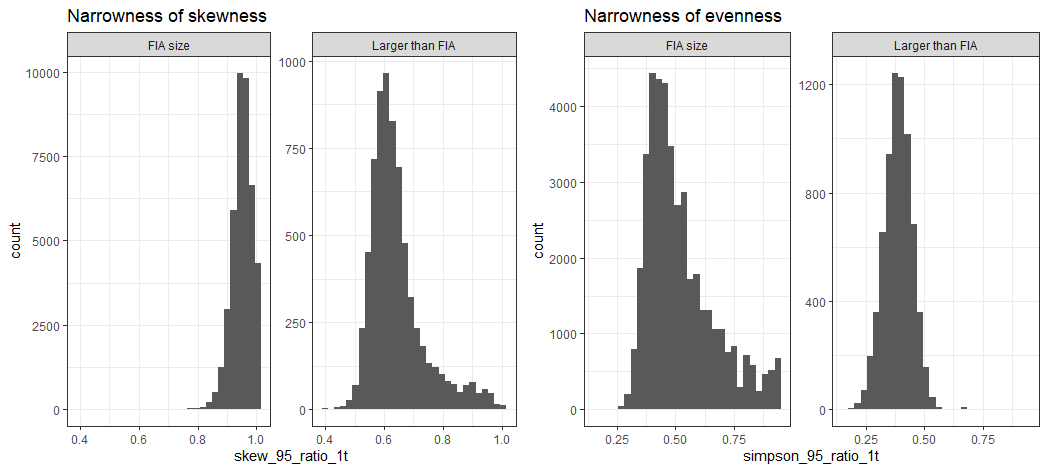


Figure 7: Rarefaction change in percentile value

