1

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

2 **Running title:** SADs deviate from statistical baselines

3 **Author names and affiliations:**

4 Renata M. Diaz*[1], Hao Ye[2], S. K. Morgan Ernest[3]

5 [1] School of Natural Resources and Environment, University of Florida, Gainesville, Florida, USA.
6 renata.diaz@weecology.org; *corresponding author

7 [2] Health Science Center Libraries, University of Florida, Gainesville, Florida, USA. haoye@ufl.edu

8 [3] Department of Wildlife Ecology and Conservation, University of Florida, Gainesville, Florida, USA.
9 skmorgane@ufl.edu

10 **Address for correspondence:**

11 Renata M. Diaz
12 110 Newins-Ziegler Hall
13 PO Box 110430,
14 Gainesville, FL 32611-0430
15 renata.diaz@weecology.org
16 Phone: (352) 846-0643
17 Fax: (352) 392-6984

18 **Statement of authorship:** RMD and SKME conceived the analysis; HY devised the algorithm to sample the
19 feasible set, reviewed the coded implementation, and wrote the explanatory vignette; RMD conducted
20 the analyses and wrote the first draft of the manuscript; all authors contributed substantively to revisions.

21 **Data accessibility statement**: All data used are available publicly via Zenodo and figshare. The data used
22 are available at https://doi.org/10.6084/m9.figshare.c.3304845.v1 and
23 http://dx.doi.org/10.6084/m9.figshare.95843. The main codebase for these analyses, including all data,
24 can be accessed at https://doi.org/10.5281/zenodo.4711104, while the R package for sampling from the
25 feasible set is at https://doi.org/10.5281/zenodo.4710750.

26 **Keywords:** Species abundance distributions; feasible set; combinatorics; macroecology; constraints

27 **Conflict of interests:** The authors declare no conflicts of interest.

28 **Type of article**: Synthesis

29 **Word counts:** Abstract: 199; main text: 7497; no text boxes

30 **Number of references:** 53

31 **Number of figures, tables, and text boxes:** 6 figures; 1 table; 0 text boxes

32

33   Abstract

34   Exploring and accounting for the emergent properties of ecosystems as complex systems is a promising

35   horizon in the search for general processes to explain common ecological patterns. For example, the

36   ubiquitous hollow-curve form of the species abundance distribution is frequently assumed to reflect

37   ecological processes structuring communities, but can also emerge as a statistical phenomenon from the

38   mathematical definition of an abundance distribution. Although the hollow curve may be a statistical

39   artefact, ecological processes may induce subtle deviations between empirical species abundance

40   distributions and their statistically most probable forms. These deviations may reflect biological processes

41   operating on top of mathematical constraints and provide new avenues for advancing ecological theory.

42   Examining ~22,000 communities, we found that empirical SADs are highly uneven and dominated by rare

43   species compared to their statistical baselines. Efforts to detect deviations may be less informative in

44   small communities – those with few species or individuals – because these communities have poorly-

45   resolved statistical baselines. The uneven nature of many empirical SADs demonstrates a path forward for

46   leveraging complexity to understand ecological processes governing the distribution of abundance, while

47   the issues posed by small communities illustrate the limitations of using this approach to study ecological

48   patterns in small samples.

49   Introduction

50   Ecological communities are complex systems made of numerous interacting entities subject to a vast

51   array of processes operating in different contexts and at different scales (Levin 1992; Lawton 1999;

52   Maurer 1999; Brown et al 2002; Nekola and Brown 2007; McGill 2019). One strategy for making sense of

53   this inherent complexity is to identify patterns that occur consistently across many communities, and use

54   these common phenomena to develop and test theories regarding general mechanisms that shape

55   community structure (Brown and Maurer 1989; Maurer 1999; Lawton 1999; Gaston and Blackburn 2000;

56   McGill 2019). Some of these patterns, however, can have counterintuitive emergent statistical properties

57   (Frank 2009; 2019). Left unexamined, these properties can confound the interpretation of the observed

58   patterns: what we interpret to be the result of generative mechanism may be an artifact of statistical

59   constraints. However, when these properties are properly understood and accounted for, they can

60   provide leverage for detecting and identifying the processes at work in a system (Jaynes 1957, Harte and

61   Newman 2014).

62   The species abundance distribution (SAD) – the distribution of how all of the individuals in a community

63   are divided among the species in that community – is a prime example of an ecological pattern that is

64   both commonly invoked in the search for general processes, and subject to statistical constraints that

65   have thus far complicated efforts to use it in this way (Nekola and Brown 2007; McGill et al. 2007; Locey

66   and White 2013). The shape of the SAD is so consistent that it is often considered an ecological law

67   (Preston 1948, 1962a, 1962b, 1980; Lawton 1999, McGill 2003, McGill et al. 2007). Across varied

68   ecosystems and taxa, the species abundance distribution is dominated by a few very abundant species

69   and a larger number of increasingly rare species, generating a distinctive hollow- or J-shaped curve when

70   plotted with species rank on the x-axis and abundance on the y-axis (Fisher et al. 1943; McGill et al 2007).

71   Community ecologists have used the SAD to test numerous theories regarding which biological processes

72   are most important for structuring assemblages of species, by comparing theoretical predictions for the

73   SAD to observed SADs (McGill 2003; McGill et al. 2007). However, this approach has proven inconclusive

74   because many theories predict similar shapes for the SAD (McGill 2003; McGill et al. 2007), and even

75   experimental manipulations generate little variation in the shape of the SAD (Supp and Ernest 2014).

76   Investigating and accounting for the statistical considerations that constrain the shape of the SAD may

77   open up new avenues for ecological interpretations of the SAD.

78   In fact, the nearly ubiquitous shape of the SAD may transcend ecological processes and instead reflect

79   mathematical properties inherent to abundance distributions. Complex systems across domains ranging

80   from economics to information technology often exhibit empirical abundance distributions with hollow-

81   curve forms similar to ecological SADs (Shockley 1957; Gaston et al. 1993; Nekola and Brown 2007,

82   Blonder et al. 2014; Keil et al. 2018). This suggests that the hollow curve is a common feature of

83   abundance distributions and not necessarily an ecological phenomenon. Because the hollow-curve is

84   observed in diverse systems and many theoretical generative processes converge to power-law or log-

85   series abundance distributions (i.e. hollow curves) (Preston 1950; McGill 2003; Nekola and Brown 2007;

86   Frank 2009; Frank 2019), approaches from statistical mechanics and complexity science may best explain

87   the expected emergent shape for the distribution (Preston 1950; McGill 2003; Nekola and Brown 2007;

88   Dewar and Porté 2008). Indeed, frameworks grounded in both entropy maximization (e.g. the Maximum

89   Entropy Theory of Ecology; Harte et al. 2008, Harte 2011) and combinatorics (i.e. 'the feasible set'; Locey

90   and White 2013) generate realistic hollow curves via the random division of the total number of

91   individuals in a community, N, into the total number of species present S. If the SAD is statistically inclined

92   to be a hollow curve, the hollow-curve in itself may be of limited use for developing and testing ecological

93   theories.

94   While SADs may be statistically constrained, this does not necessarily mean that they cannot be

95   biologically informative. Biological factors may introduce subtle, but meaningful, deviations between

96   observed SADs and the shapes of the SADs expected due to the mathematical constraints imposed by S

97   and *N,* which we hereafter refer to as the "statistical baseline" (Locey and White 2013, Harte and

98   Newman 2014). If the vast majority of mathematically achievable SADs for a community share a similar

99   shape, an empirically observed SAD that deviates even slightly from this statistical baseline is unlikely to

100   have emerged at random (Locey and White 2013), and may be the signature of a non-random – i.e.,

101   biological – process operating on the relative abundances of species (Harte and Newman 2014). If, over

102   many communities, there are consistent deviations between observed SADs and their statistical

103   baselines, these deviations can help evaluate and refine ecological theories. For example, the high

104   prevalence of rare species in ecological communities has attracted considerable empirical and theoretical

105   attention (e.g. Nee et al. 1991; Magurran and Henderson 2003), but it is unclear to what extent this

106   phenomenon may derive from mathematical constraints on the SAD rather than ecological processes. If

107   the prevalence of rare species in observed distributions consistently exceeds what would be expected to

108   emerge from the statistical baseline, we would be prompted to look for ecological mechanisms

109   promoting rarity. Candidate theories could then be evaluated based on how well their predictions for the

110   rare tail of the SAD matched observed distributions. Thus, the *deviations* from the statistical baseline may

111   enable us to detect strong ecological processes or evaluate theories (Harte and Newman 2014, Xiao et al.

112   2016).

113   Successfully interpreting SADs in this fashion depends on our capacity to detect and quantify deviations

114   between empirical observations and statistical baselines, which requires metrics and computational

115   approaches that allow us to quantify and interpret whatever deviations may exist. Here, we build upon

116   the combinatoric approach developed by Locey and White (2013) to define and explore statistical

117   baselines for SADs. For a given *N* (total number of individuals) and *S* (total number of species), there exists

118   a finite set of possible distributions of individuals into species. Collectively, this set of possible SADs is the

119   *feasible set*, with each possible SAD constituting a single element of the set. If an observed SAD is drawn

120   at random from the feasible set, it is likely to have a shape similar to the shapes most common in the

121   feasible set. The feasible set therefore allows us to define statistical baselines for assessing deviations

122   between observed SADs and what is likely to occur due to mathematical constraints (Locey and White

123   2013).

124   The feasible set can also be used to explore how the characteristics of the statistical baseline, and the

125   presence and nature of any deviations that occur, vary over ranges of values for S and N. Although most

126   feasible sets are dominated by the hollow-curve shape, variation in S, N, and the ratio of N to S modulate

127   the detailed attributes of the SADs in a feasible set (Locey and White 2013). For example, if the ratio of N

128   to S is close to 1, all possible SADs are mathematically constrained to be fairly even (Locey and White

129   2013). Although an SAD that is very even would be highly unusual in most cases, it would be expected in

130   this situation. The feasible set therefore allows us to appropriately calibrate our expectations for what

131   types of observations would be surprising for an SAD given the specific constraints imposed by its S and N.

132   Additionally, accounting for variation in the specificity, or vagueness, of the expectations derived from the

133   statistical baseline may be critically important for disentangling the aspects of the SAD that can be

134   attributed to statistical constraints from those that result from other processes. If the vast majority of

135   mathematically possible SADs are similar in shape – generating a very specific, narrowly defined statistical

136   baseline – then even small deviations between an observed SAD and this baseline can signal the

137   operation of ecological processes. However, if many different shapes occur with more even frequency in

138   the feasible set, the statistical baseline is less specific and less well defined, and our sensitivity for

139   distinguishing biological signal from statistical constraints is greatly reduced. This is more likely to occur

140   when the size of the community, in terms of S and N, is small, because in such cases the feasible set may

141   be too small for a particular shape to emerge as the most common shape. These statistical baselines with

142   broad distributions may therefore impede our ability to assess whether observed deviations are

143   ecologically generated or expected to emerge randomly (Jaynes 1957). This general concern has been

144   acknowledged in efforts to compare ecological observations to statistical baselines (Harte 2011, White et

145   al. 2012, Locey and White 2013) but there has not yet been a quantification of these effects for the SAD

146   or an identification of the range of community sizes most strongly affected. Because ecologists study the

147   SAD for communities varying in size from the very small – S and $N < 5$ – to the enormous – S and $N \gg$

148   1000 – identifying the community sizes for which we can and cannot confidently detect deviations from

149   the statistical baseline is necessary to appropriately contextualize our interpretations.

150   Here we use the feasible set to define statistical baselines for empirical SADs for 22,000 communities of

151   birds, mammals, trees, and miscellaneous other taxa. We then compare *observed* SADs to their

152   corresponding statistical baselines and evaluate 1) if the shapes of observed SADs consistently deviate

153   from their statistical baseline, 2) how the characteristics and specificity of the statistical baseline vary

154   over ranges of S and N, and 3) whether this variation appears to be associated with variation in our

155   capacity to detect deviations between observations and the corresponding baselines.

156   **Methods**

157   Data and code for all of our analyses can be accessed at http://doi.org/10.5281/zenodo.4711104.

158   *Datasets*

159   We used a compilation of community abundance data for trees, birds, mammals, and miscellaneous

160   additional taxa (White et al. 2012, Baldridge 2015, Baldridge 2016, data from Baldridge 2016). This

161   compilation consists of cleaned and summarized community abundance data for trees obtained from the

162   Forest Inventory and Analysis (Woudenberg et al 2010) and Gentry transects (Phillipes and Miller 2002),

163   birds from the North American Breeding Bird Survey (Sauer et al. 2013), mammals from the Mammal

164   Community Abundance Database (Thibault et al. 2011), and a variety of less commonly sampled taxa

165   from the Miscellaneous Abundance Database (Baldridge 2015). Because characterizing the random

166   expectation of the SAD is computationally intractable for very large communities, we filtered our datasets

167   to remove 4 communities that had more than 40714 individuals, which was the largest community we

168   successfully analyzed. We further filtered the FIA database. Of the 103,343 communities in FIA, 92,988

169   have fewer than 10 species. Rather than analyze all these small communities, we randomly selected

170   10,000 small communities to include in the analysis. We also included all FIA communities with more than

171   10 species, which added 10,355 FIA communities to the analysis and resulted in a total of 20,355 FIA

172   communities. Finally, for sites that had repeated sampling over time, we followed White et al. (2012) and

173   Baldridge (2016) and analyzed only a single, randomly selected, year of data, because samples taken from

174   a single community at different time points are likely to covary. It should be noted that our analyses

175   include data from the Mammal Community Database and Miscellaneous Abundance Database that were

176   collected over longer timescales and cannot be disaggregated into finer units of time. Our final dataset

177   consisted of ~22,000 communities with S and N ranging from 2 to 250 and 4 to 40714, respectively

178   (Figure 1). Details and code for the filtering process can be found in Appendix S1 in Supporting

179   Information.

180   *Accounting for empirical sampling error*

181   Because it is logistically impossible to exhaustively census all individuals present in most empirical

182   systems, SADs derived from field sampling will inevitably be subject to some degree of sampling error

183   (Bonar et al. 2011). Therefore, in addition to analyzing the raw SADs in our database, we employed two

184   resampling schemes to test if, and how, different forms of observation error affect our results.

185   First, we explored the possibility that empirical sampling systematically undercounts the true number of

186   rare species in a community (Preston 1948; Gotelli and Colwell 2011). Rare species are more likely to

187   escape detection during sampling, leading to an underestimate of both the total species richness of a

188   community and the proportion of species in the rare tail of the SAD (Preston 1948). We used a procedure

189   based on species richness estimators to adjust for this possibility (see also Ulrich et al. 2010 for the use of

190   richness estimators to distinguish between completely and incompletely censused communities). We

191   computed the estimated richness for each community using the bias-corrected Chao and the ACE

192   estimators (as implemented in the R package "vegan"; O'Hara 2005; Chiu et al 2014; Oksanen et al.

193   2020). To each of these richness estimates, we added one standard deviation of the estimate, and then

194   took the mean of the two results. This yields a generous estimate of the true number of species in the

195   system. If this estimate exceeded the observed species richness, we added the missing species each with

196   abundance 1. These adjusted SADs allowed us to explore the consequences of undersampling rare

197   species while making the smallest possible changes to S and N.

198   Second, we tested the sensitivity of our results to sampling variability across all species in the SAD – not

199   just rare species - using subsampling. For each observed community, we constructed subsamples by

200   randomly drawing 60% of the observed number of individuals from the total pool of individuals in the

201   community, without regard to species and without replacement. The precise proportion of individuals

202   drawn in each subsample should not dramatically affect the qualitative outcome. We selected 60% so as

203   to introduce appreciable room for sampling error between the raw and subsampled SADs, but to produce

204   subsampled SADs with N (and presumably S) in a comparable size range to the raw ones. Extremely small

205   subsamples (e.g. 10%) could introduce complications related to small N and S that could obscure the

206   effects of sampling error, while very large subsamples (e.g. 90%) could recapture the raw distributions

207   too closely to be informative. We generated 10 resampled communities for each observed community.

208   We ran our computational pipeline using all raw SADs and all SADs adjusted for undersampling of rare

209   species. Because the subsampling approach increased computational effort approximately tenfold, we

210   analyzed all subsampled communities for the Mammal Community, Miscellaneous Abundance, and

211   Gentry databases, but only a random subset of 300 (of 2773) communities from the Breeding Bird Survey

212   and 2000 (of 20179) from the FIA – 1,000 with S < 10, and 1,000 with S >= 10.

213   *Generating the statistical baseline*

214   We use the concept of the "feasible set" to establish a statistical baseline for the SAD (Locey and White

215   2013). For a given number of individuals N, there are a finite number of unique ways to partition those

216   individuals into S species. The complete set of these unique partitions is the feasible set. In this approach,

217   neither species nor individuals are distinguishable from each other; thus partitions are unique if and only

218   if they differ in the number of species that have a particular abundance (Locey and White, 2013).

219   Operationally, this means that for $S = 3$ and $N = 9$, the SADs *(1, 3, 5)* and *(2, 2, 5)* count as distinct

220   partitions, but *(1, 3, 5)* and *(3, 1, 5)* do not, because they each contain one species with an abundance 1,

221   3, and 5, respectively, and differ only in the *order* of the numbers. In the absence of justification for

222   additional assumptions regarding the distinguishability of species and/or individuals, we adopted this

223   simple set of assumptions that has previously been shown to generate realistic statistical baselines (Locey

224   and White 2013).

225   While it is possible to list all possible partitions in the feasible set for small S and N, the size of the feasible

226   set increases rapidly with S and N. An exhaustive characterization of the statistical properties of the

227   feasible set for large S and N quickly becomes computationally intractable. This renders it necessary to

228   draw samples from the feasible set, rather than enumerating all of its elements. Previous efforts in this

229   vein (Locey and White 2013) have been constrained by the problem of unbiased sampling of large

230   feasible sets. We developed an algorithm to efficiently and uniformly sample feasible sets even for large

231   values of S and N. In brief, the algorithm takes a generative approach to sample the feasible set for a

232   given combination of S and N, based on recurrence relations used to calculate the size of the feasible set.

233   Let f(S, N) be the number of possible partitions of N individuals into exactly S species, i.e. the size of the

234   feasible set for given values of S and N. Computation of f(S, N) can be achieved without enumerating the

235   entire feasible set through the recurrence relation f(S, N) = f(S-1, N-1) + f(S, N-S) (originally documented

236   in a 1742 letter from Euler to Bernoulli; 1862). For example, consider the feasible set with S = 3 and N = 7.

237   For all possible partitions, either (a) at least one species has an abundance equal to 1, or (b) all of the

238   species have abundance greater than 1. In the case of (a), removing one species with abundance equal to

239   1 must result in a partition of 6 individuals into 2 species. In fact, all of the unique partitions in (a) must

240   have a corresponding unique partition in the feasible set for S = 2 and N = 6, and vice versa. In the case of

241   (b), removing 1 individual from each species must result in a partition from the feasible set with S = 3 and

242   N = 4. Here, all the partitions in (b) must have a corresponding unique partition in the feasible set with S =

243   3 and N = 4, and vice-versa. Therefore, f(3,7) = f(2,6) + f(3,4). By storing the values in a lookup table, f(S,

244   N) can be calculated for increasing values of S and N through straightforward summation.

245   This recurrence relation also makes it possible to draw random samples from the feasible set without

246   enumerating all possible partitions of N into S. For the example of S = 3 and N = 7, there are a total of 4

247   possible partitions (i.e. f(S, N) = 4). Because f(2, 6) = 3 and f(3, 4) = 1, we know that (a) 3 of the 4

248   partitions must correspond to a partition of the feasible set with S = 2 and N = 6 (but with a species of

249   abundance equal to 1 removed), and (b) 1 of the 4 partitions must correspond to a partition of the

250   feasible set with S = 3 and N = 4 (but with 1 individual removed from each species). Thus, we can

251   determine the probability that a partition drawn at random from the feasible set for S = 3 and N = 4 is in

252   case (a) – probability ¾ - or case (b) – probability ¼. To generate a partition in case (a), we sample a

253   partition for S = 2 and N = 6 and then add a species with abundance equal to 1; for case (b), we sample a

254   partition for S = 3 and N = 4 and then add 1 individual to each species. In this way, we use the recurrence

255   relation to transform the problem of sampling from a large feasible set into the problem of sampling from

256   a smaller, different feasible set. This procedure continues until a partition is uniquely determined, after

257   which some back-transformation yields a unique partition for the feasible set of interest. A detailed

258   description of the algorithm we use, based on a slightly different recurrence relation, is available in

259   Appendix S2 and is implemented in the R package `feasiblesads` available at

260   http://doi.org/10.5281/zenodo.4710750.

261   For every community in our database, we drew 4000 samples from the feasible set to characterize the

262   distribution of statistically probable shapes for the SAD. We filtered the 4000 samples to unique

263   elements. For small values of S and N, it can be impossible or highly improbable for the 4000 samples

264   from the feasible set to all be unique, but for large communities, all 4000 are usually unique. We refer to

265   this as the sampled feasible set.

266   *Comparing observed SADs to their statistical baselines*

267   We compared SADs to their statistical baselines using several metrics, including a general measure of

268   dissimilarity, as well as skewness, Simpson's evenness, Shannon's index, and the proportion of rare

269   species (species with abundance = 1). These metrics represent just a few of the vast array of possible

270   summary metrics to describe the shape of the SAD, each of which emphasize different aspects of the

271   distribution. In this first effort to compare empirical distributions to a statistical baseline, we selected a

272   suite of complementary metrics and explored whether our overall results were consistent between

273   metrics. By calculating these metrics for each the community's sampled feasible set (see *Generating the*

274   *statistical baseline*, above), we generated a portfolio of measures describing the shapes expected from

275   randomly sampled SADs.

276   First, as a general characterization of whether observed SADs have rare or common shapes relative to

277   their feasible sets, we computed a dissimilarity score comparing SADs to the central tendencies of their

278   feasible sets (following Locey and White, 2013). We defined the degree of dissimilarity between two SADs

279   with the same S and N as the proportion of individuals allocated to species with different abundances

280   between the two SADs, calculated as:

281   $1 - \frac{\sum_{i=1}^{S} |n1_i - n2_i|}{2N}$

282   where $n1_i$ is the abundance at rank i for one SAD and $n2_i$ is the abundance at rank i for the other SAD. This

283   value ranges from 0 to 1, with 1 being high dissimilarity. To find the central tendency of a given sampled

284   feasible set, we identified the sampled SAD with the lowest mean dissimilarity compared to the rest of

285   the SADs in the feasible set. We calculated the dissimilarity between every sample drawn from the

286   feasible set and a random set of 500 other samples, using a subset of samples for comparisons because it

287   is computationally impractical to make all pairwise comparisons between large numbers of samples. To

288   assess whether an observed SAD was highly dissimilar to its central tendency, we calculated the degree of

289   dissimilarity between the central tendency of the corresponding feasible set and all other samples from

290   that feasible set, and between the central tendency and the observed SAD. Although the dissimilarity

291   score is scaled from 0 to 1, the distributions of dissimilarity scores for samples from the feasible set can

292   vary over broad ranges in S and N. We therefore used the percentile rank of the observed dissimilarity

293   scores, relative to the distribution of dissimilarity scores from the corresponding sampled feasible sets, to

294   quantify how likely or unlikely observed dissimilarity scores are across the range of S and N in our

295   datasets. For a single community, an observed percentile score of 95 indicates that there is a 5% chance

296   of drawing a value greater than the observed value from the distribution of values from the sampled

297   feasible set. Aggregating across communities, if observed SADs reflect random draws from their feasible

298   sets, their percentile rank values should be uniformly distributed from 0 to 100. However, if observed

299   SADs are consistently more dissimilar to their feasible sets that expected at random, the percentile values

300   will be disproportionately concentrated at high values. We used a one-tailed 95 confidence interval and

301   tested whether the percentile values for the dissimilarity scores of observed SADs fell above 95 more

302   than 5% of the time. We note that it is impossible for an observation fall above the 95th percentile if there

303   are fewer than 20 values in the sampled distribution. We therefore excluded from this analysis

304   communities with fewer than 20 unique SADs in their feasible sets, yielding a total of 22,490

305   communities. Finally, note that, if the observed dissimilarity scores for individual communities are not

306   systematically higher than the distributions of dissimilarity scores from the corresponding feasible sets,

307   increasing the number of *communities* in the analysis will not increase the frequency of extreme

308   percentile scores.

309   While the degree of dissimilarity between SADs and the central tendency of the feasible set provides an

310   overall sense of deviations among possible SADs, it does not describe *how* observed SADs may differ from

311   their feasible set. We therefore used a set of more targeted, ecologically interpretable metrics to explore

312   how observed SADs compare to their feasible sets in their shape and proportion of rare species. We

313   examined three metrics for the shape of the SAD - skewness, Simpson's evenness (1-D), and Shannon's

314   index. Skewness measures the asymmetry of a distribution around its mean. The Simpson and Shannon

315   indices are commonly used metrics for assessing how equitably abundance is distributed across species

316   (Maurer and McGill 2011). We also calculated the proportion of rare species (species with abundance = 1)

317   in each SAD, because the proportion of rare species in a community is comparable across different

318   community sizes and is of special interest to ecologists.

319   As with the degree of dissimilarity score, to assess whether the shape of an observed SAD was statistically

320   unlikely, we used percentile ranks to compare the observed values of the summary metrics to the

321   distributions of values for those metrics obtained from each community's sampled feasible set. The actual

322   ranges and values of summary metrics vary widely over the ranges of S and N in our data and thus cannot

323   directly compared, but percentile ranks are comparable across different community sizes and allow

324   assessment across our entire dataset. We used two-tailed 95% intervals to test whether observed

325   communities' percentile values for each metric were disproportionately concentrated below 2.5 or above

326   97.5. In all cases, in testing for unusually high percentile scores, we defined the percentile score as the

327   proportion of values in the sampled distribution strictly less than the observed value, while in testing for

328   low values, we defined it as the proportion of sampled values less than or equal to the observed value.

329   This ensured a conservative estimate of how extreme the observed values were relative to the sampled

330   distribution. Because it is impossible for an observed percentile score to be above or below the 97.5th or

331   2.5th percentile if there are fewer than 40 values in the sample distribution, we excluded from these

332   analyses communities with fewer than 40 SADs in their feasible sets. Finally, note that skewness, as

333   implemented in the R package "e1071" (Meyer et al. 2019), always evaluates to 0 for distributions with

334   only two species, and we therefore excluded those cases from analyses of skewness. Our final analysis

335   included 21,395 communities for skewness and 21,403 communities for all other shape metrics.

336   *The narrowness of the expectation*

337   We also used the distributions of dissimilarity scores and shape metrics to quantify the relative specificity

338   of the statistical baseline, in order to assess when there could be challenges in determining whether

339   observed communities differ from their statistical baselines. For an overall sense of how tightly elements

340   of the feasible set were clustered around its central tendency, we calculated the 95th percentile of

341   dissimilarity scores for all samples from a feasible set compared to the central tendency of that feasible

342   set. This value shows how dissimilar to the central tendency an observed SAD would need to be to count

343   as "extreme" relative to the feasible set. For the shape metrics, we calculated a breadth index defined as

344   the ratio of the range of values encompassed within a two-sided 95% density interval relative to the full

345   range of values in the distribution (Figure 2). This breadth index for the statistical baseline ranges from 0

346   (a very narrow distribution and well-resolved baseline) to 1 (a very broad distribution), and is comparable

347   across feasible sets for varying combinations of *S* and *N*. These approaches correspond qualitatively to

348   more computationally-intensive approaches to measuring the self-similarity of the elements of feasible

349   sets (see Appendix S3). We explored how the narrowness of the statistical baseline varies with the size of

350   the feasible set and the ratio of N to S.

351   **Results**

352   *Comparing observed SADs to their statistical baselines*

353   For four of the five datasets we analyzed – BBS, Gentry, Mammal Communities, and Miscellaneous

354   Abundance – observed SADs are more dissimilar to their statistical baselines than would be expected by

355   chance (Figure 3). Combined over these four datasets, 29% of observed SADs are more dissimilar to the

356   central tendency than are 95% of samples from the corresponding feasible sets (Table 1). If observed

357   SADs reflected random draws from the feasible set, we would expect only 5% to be that dissimilar. These

358   highly unlikely SADs have dissimilarity scores from 1.007 to 6.3 times the 95th percentile of dissimilarity

359   scores comparing the central tendency to samples from the feasible set, an absolute increase ranging

360   from .001 to .54 on a scale from 0-1 (Figure S4). These datasets also contain highly unlikely observed

361   SADs in terms of their shape metrics. At random, roughly 2.5% of observed percentile scores for these

362   metrics should be very high (>97.5) or very low (<2.5). Compared to their feasible sets, these four

363   datasets contain a disproportionate number of communities with very low values for Simpson's evenness

364   and Shannon diversity, and very high skewness, relative to their feasible sets (Table 1). The Mammal

365   Community and Miscellaneous Abundance databases also have high proportions of rare species, but this

366   tendency is weaker for BBS and nonexistent for Gentry – in fact, the Gentry dataset has a high

367   representation of sites with *low* proportions of rare species (20% of sites; Table S5). The Gentry dataset

368   also has a disproportionate number of communities with the opposite tendencies to the other datasets

369   for the other shape metrics– i.e., an overrepresentation of communities with high Simpson's evenness

370   and Shannon diversity, and low skewness.

371   In contrast to the other datasets, percentile scores for sites from the FIA dataset are more uniformly

372   distributed, and the proportions of extreme values are closer to what would be expected by chance

373   (Figure 3, Table 1). Only 7% of FIA communities are highly dissimilar to their feasible sets (compared to a

374   random expectation of 5%). Among the shape metrics, only 2.8% (compared to 2.5% at random) of sites

375   have high values for skewness, 1.3% have high proportions of rare species, 5.7% have low Simpson's

376   evenness, and 5.4% have low Shannon diversity.

377   *The narrowness of the expectation*

378   The ability to detect deviations from the statistical baseline depends in part on the distribution of SADs in

379   the feasible set. Overall, as the size of the feasible set increases, the SADs in a feasible set become more

380   narrowly clustered around the central tendency of that feasible set, and the sampled distributions for

381   shape metrics generally become less variable (Figure 4). In small communities, the breadth indices are

382   highly variable and often very large – approaching 1, meaning that a 95% density interval of the values in

383   the distribution spans nearly the entire range of values – while the breadth indices for larger communities

384   rarely exceed ~.7 for skewness, Simpson evenness, and Shannon diversity, and ~.8 for the proportion of

385   rare species. Among our datasets, the FIA and Mammal Community databases have the smallest

386   communities, in terms of S and N, and tend to have the largest proportions of feasible sets with high

387   breadth indices (Figure S6).

388   *Sensitivity to sampling variability*

389   In almost all cases, SADs adjusted for the under-observation of rare species are even more extreme

390   relative to their feasible sets than unadjusted SADs (Figure 5; see Appendix A7 for complete results of

391   resampling). For all datasets, adjusted SADs show more high values for skewness and the proportion of

392   rare species, and low values for Simpson's evenness and Shannon diversity, than unadjusted SADs.

393   Subsampling consistently reduces the proportion of extreme observations across all datasets and metrics

394   (Figure 5; Appendix A7). In most instances, the proportion of extreme observations still exceeds the

395   proportion that would be expected by chance. However, the proportion of sites with high numbers of

396   rare species observed for the BBS and Mammal Community databases drop from 4.5% to 1% and ~13% to

397   3.5% with resampling. For FIA, the proportions of sites with high dissimilarity, low evenness and Shannon

398   diversity all drop from 6-8% to 2-3%. Note that, for FIA, neither the raw nor the resampled SADs have a

399   disproportionate representation of extreme values for the remaining metrics.

400   Discussion

401   We found widespread evidence that SADs for a range of real ecological communities deviate from the

402   forms expected given the distribution of shapes within their feasible sets. Overall, these deviations may

403   signal that ecological processes operate on top of statistical constraints, thereby driving the SAD away

404   from shapes generated by purely statistical processes. We also found that the magnitude and form of

405   deviation varied among the datasets we considered. This variability may reflect statistical phenomena

406   related to the size of S and N and their ratio, or it may reflect different biological processes dominating in

407   different contexts. Finally, although a disproportionate number of communities deviated statistically from

408   their feasible sets, there were also many communities for which we did not detect deviations. This does

409   not imply the absence of ecological processes operating on these SADs. Rather, one possible explanation

410   is that multiple ecological processes are operating simultaneously and with countervailing effects,

411   resulting in no dominating net impact on the shape of the distribution beyond that imposed by

412   fundamental constraints (Harte 2011; Harte and Newman 2014). Going forward, testing whether

413   ecological theories or common functional approximations (e.g. the log-normal distribution) accurately

414   predict the deviations between observed SADs and their statistical baselines may be much more fruitful

415   than focusing only on the general form of the SAD (McGill et al. 2007; Locey and White 2013; Harte and

416   Newman 2014).

417   In most cases, and most pronouncedly for the Breeding Bird Survey, Mammal Community, and

418   Miscellaneous Abundance databases, our results suggest that the prevailing processes cause abundance

419   distributions to be highly uneven, rather than those that produce more even abundances across species.

420   For these communities, observed SADs tended to be unusually skewed and uneven, and to have a high

421   proportion of rare species, compared to their feasible sets. Accounting for undersampling of rare species

422   strengthened these effects, while subsampling weakened them. Perhaps unsurprisingly, the effect of

423   these two resampling approaches was especially noticeable for the proportion of rare species; enriching

424   the SAD directly adds rare species, while subsampling is likely to drop rare species even if it otherwise

425   recaptures the general shape of a distribution. The long tail of rare species in the SAD has been a

426   consistent focus in SAD research, and our results highlight that the rare tails of observed SADs are

427   extraordinary, even among the hollow-curve shapes that dominate the feasible set. Ecological processes

428   may lengthen the rare tail and decrease the evenness of the SAD, for example by promoting the

429   persistence of rare species at very low abundances (Yenni et al. 2012). Or, they could drive abundant

430   species to have larger populations than would be statistically expected, without also driving other species

431   entirely to extinction (Chesson 2000).

432   While the Gentry database also exhibits deviations tending towards high unevenness, an even greater

433   proportion of its communities are *more* even, and have a lower proportion of rare species, than would be

434   expected given their feasible sets. This could indicate that there are biological differences between the

435   systems in the Gentry and other datasets that result in different forms for the SAD. Alternatively, the

436   statistical characteristics of the feasible sets for these communities could modulate the detected

437   deviations. Communities in the Gentry database have high species richness and low average abundance

438   (Figure 1). Among these, many of the communities exhibiting high evenness and low proportions of rare

439   species are those with very high species richness and low average abundance (N/S < ~3) (see Appendix

440   A8). As a result, these communities have unusual statistical baselines: for example, the corresponding

441   feasible sets have the highest proportions of rare species of any of the feasible sets in our analysis.

442   Although observed SADs for these communities also have high proportions of rare species, taking the

443   statistical baseline into account would suggest that the extraordinary thing about these SADs is that they

444   do not have even more rare species. Simultaneously, there may be biological reasons why the species-

445   rich but relatively low-abundance tropical tree communities of the Gentry database differ from those in

446   other datasets. The same mechanisms that promote high diversity may manifest in high evenness, and/or

447   ecological features particular to these forests may produce unusual shapes for the SAD. Because no

448   communities from our other datasets are comparable in S and N, we cannot disentangle these statistical

449   and biological explanations. This is an excellent opportunity to develop additional theoretical and

450   empirical approaches to predict and explain variation in the deviations between SADs and their feasible

451   sets, in particular for species-rich communities across ecosystems.

452   Unlike the other four datasets, communities in the FIA dataset showed weak or no evidence of deviations

453   from their feasible sets. We entertained two general classes of explanation for why the FIA dataset differs

454   from the others in our analysis: first, that biological attributes of the FIA communities cause the SADs for

455   these communities to differ from the others in our database, and second, that statistical phenomena

456   related to S and N may modulate the capacity to detect deviations for these communities. To distinguish

457   between possible biological drivers causing FIA to differ from the other datasets, and factors intrinsic to S

458   and N, we compared a subset of ~300 FIA communities to communities from other datasets with directly

459   matching S and N. We did not find differences in the distribution of percentile scores for any metrics

460   between communities from FIA and communities from other datasets, confirmed via Kolmogorov-

461   Smirnov tests (Appendix A9). Although 300 communities constitute a small sample relative to the 20,355

462   FIA communities we analyzed, these results point to statistical phenomena, and not biological attributes

463   unique to FIA, as the likely explanation for the differences.

464   A second possibility is that these differences reflect statistical phenomena related to community size in

465   terms of S, N, and as a result, the number of possible SADs in a community's feasible set. The FIA

466   communities are the smallest across our datasets (Figure 1), and communities with small values of S and

467   N have smaller feasible sets. When there are relatively few possible SADs in the feasible set, they may be

468   less tightly clustered around their central tendencies, and the distributions for their shape metrics may be

469   less narrowly peaked, than when there are very large numbers of possible SADs. High variability within

470   the feasible set weakens the statistical distinction between "common" and "extreme" shapes (Figure 2).

471   Under these circumstances, any deviations – or lack thereof – will be less informative than for

472   communities with more strongly defined statistical baselines (Jaynes 1957). The dissimilarity to the

473   central tendency, and the distributions of breath indices for specific metrics, broadly align with this

474   principle. Across the range of community sizes represented in our datasets, small feasible sets have highly

475   variable, and often very broad, feasible sets (Figure 4). More specifically, very small communities – for

476   example, those with fewer than 2000 possible SADs in their feasible sets, or S ~20 and N ~40 – exhibit

477   more highly variable feasible sets than large communities, and these small communities also show less

478   consistent deviations (Figure 6; Appendix A10). Of our datasets, FIA is most dominated by small

479   communities (68% of communities have fewer than 2000 possible SADs), and these small-community

480   phenomena may therefore have the greatest impact on results aggregated over the FIA dataset.

481   If it is true that the highly variable feasible sets associated with small communities contribute to the weak

482   evidence of deviations observed for the FIA dataset, such considerations affect our capacity to use this

483   approach to distinguish signal from noise for a substantial contingent of ecological communities. Because

484   the combinations of S and N represented in our analyses are irregularly distributed among different

485   datasets (Figure 1), and because there is a great deal of variation in our breadth indices not accounted for

486   by the size of the feasible set (Figure 4), we do not interpret these results as showing a threshold for

487   defining problematically small communities. A more systematic exploration of the S and N state space,

488   combined with more nuanced metrics for characterizing the variability of the feasible set, could clarify the

489   relationship between S and N, the size of the feasible set, and statistical power. However, FIA and other

490   small, highly variable communities have on the order of 10-20 species and 30-60 individuals, suggesting a

491   general range of values below which we have diminished power to detect deviations from the statistical

492   baseline represented by the feasible set. Communities with on the order of 5 species, or 100s to 1000s of

493   individuals, have previously been suggested as "small" in this context (Preston 1948; McGill et al. 2007).

494   To meaningfully draw inferences using deviations in these small communities, we will need more

495   sensitive metrics than those used here, and/or theories that generate more specific predictions for the

496   SAD. In the absence of such, we may stand to learn the most by focusing on SADs from relatively large

497   communities.

498   It is also important to recognize that there are multiple plausible approaches to defining a statistical

499   baseline for the SAD, of which we have taken only one (Haegeman and Loreau 2008, Locey and White

500   2013). Our approach follows Locey and White (2013) and reflects the random partitioning of individuals

501   into species, with the resulting distributions considered unique if the species' abundance values are

502   unique, regardless of the order in which the values occur. This philosophy reflects a longstanding

503   approach in the study of abundance distributions: to focus on the shape of the distribution without

504   regard to species' identities (McGill et al 2007). Other assumptions regarding the statistical baseline may

505   be equally valid and generate different statistical expectations, which may alter if, and in what ways,

506   empirical distributions appear unusual. For example, incorporating differences in species order into the

507   statistical baseline – which would imply that identifying *which* species contain the most or least

508   individuals is important – might reduce the representation of long-tailed, highly uneven SADs within the

509   feasible set, and make the rare tail observed for real SADs appear more unlikely than it does here. Under

510   our assumptions, the SADs (1,2,3,4) and (1, 1, 1, 7) each count as only one unique SAD. Taking species

511   order into account would mean that (1,2,3,4) would count as 24 (4!) unique SADs, because there are 4!

512   ways to assign the abundances to each species. However, an SAD containing species with equal

513   abundances, such as (1, 1, 1, 7), would only count as 4 unique SADs. For SADs, equal abundances are

514   likely most prevalent among rare species. If this is true, then this set of assumptions would generate

515   feasible sets where rare-tailed SADs are relatively scarce, making observed SADs with rare tails seem even

516   more extraordinary. Additional formulations for the statistical baseline exist, including those that

517   approximate exponential, Poisson, or log-series distributions in the limit (Harte et al. 2008, Favretti 2018).

518   Investigating and comparing the results that emerge from different baselines will be an important next

519   step towards reinvigorating the use of the SAD as a diagnostic tool.

520   Our study demonstrates the utility, and the potential challenges, of applying tools from the study of

521   complex systems and statistical mechanics to the study of ecological communities (Haegeman and Loreau

522   2008, Harte 2011, White et al. 2012, Harte and Newman 2014)**.** While concepts such as maximum

523   entropy and the feasible set are promising horizons for macroecology, the small size of some ecological

524   communities may present difficulties that are rare in the domains for which these tools were originally

525   developed (Jaynes 1957, Haegeman and Loreau 2008). When the observed numbers of species and

526   individuals are too small to generate highly resolved statistical baselines, these approaches will be less

527   informative than we might hope – as appears to be the case for the smallest communities in our analysis.

528   In larger communities, where mathematical constraints have more resolved effects on the form of the

529   SAD, our results show that these constraints alone do not fully account for the extremely uneven SADs we

530   often observe in nature – leaving an important role for ecological processes. This ability to detect and

531   diagnose the specific ways in which empirical SADs deviate from randomness can generate new avenues

532   for understanding how and when biological drivers affect the SAD. There are, of course, still many

533   elements to be improved in our ability to distinguish biological signal from randomness, including

534   assessing alternative statistical baselines and calibrating our expected power to detect deviations,

535   especially for small communities. Indeed, more sensitive metrics could also enable identification of

536   processes that operate through time. Continuing to explore and account for the interplay between

537   statistical constraint and biological process constitutes a promising and profound new approach to our

538   understanding of this familiar, yet surprisingly mysterious, ecological pattern.

539

546

547

548   References

549   Baldridge, E. (2015). Miscellaneous Abundance Database. figshare. Available at:

550   https://doi.org/10.6084/m9.figshare.95843.v4

551   Baldridge, E., Harris, D.J., Xiao, X. & White, E.P. (2016). An extensive comparison of species-abundance

552   distribution models. *PeerJ*, 4, e2823.

553   Blonder, B., Sloat, L., Enquist, B.J. & McGill, B. (2014). Separating Macroecological Pattern and Process:

554   Comparing Ecological, Economic, and Geological Systems. *PLOS ONE*, 9, e112850.

555   Bonar, S.A., Fehmi, J.S. & Mercado-Silva, N. (2011). An overview of sampling issues in species diversity and

556   abundance surveys. In: *Biological Diversity: Frontiers in Measurement and Assessment* (eds.

557   Magurran, A.E. & McGill, B.J.). Oxford University Press, Oxford, United Kingdom, pp. 11–24.

558   Brown, J.H., Gupta, V.K., Li, B.-L., Milne, B.T., Restrepo, C. & West, G.B. (2002). The fractal nature of

559   nature: power laws, ecological complexity and biodiversity. *Phil. Trans. R. Soc. Lond. B*, 357, 619–

560   626.

561   Brown, J.H. & Maurer, B.A. (1989). Macroecology: The Division of Food and Space Among Species on

562   Continents. *Science*, 243, 1145–1150.

563   Chesson, P. (2000). Mechanisms of Maintenance of Species Diversity. *Annual Review of Ecology and

564   Systematics*, 31, 343–366.

565   Chiu, C.-H., Wang, Y.-T., Walther, B.A. & Chao, A. (2014). An improved nonparametric lower bound of

566   species richness via a modified good-turing frequency formula. *Biometrics*, 70, 671–682.

567   Dewar, R.C. & Porté, A. (2008). Statistical mechanics unifies different ecological patterns. *Journal of

568   Theoretical Biology*, 251, 389–403.

569   Euler, L. (1862). Sex litterae ad Nicolaum Bernoullium II, Basileensem J. U. D. datae 1742 ad 1745. *Opera

570   Postuma*, 519–549.

571   Favretti, M. (2018). Remarks on the Maximum Entropy Principle with Application to the Maximum

572   Entropy Theory of Ecology. *Entropy*, 20, 11.

573   Fisher, R.A., Corbet, A.S. & Williams, C.B. (1943). The Relation Between the Number of Species and the

574   Number of Individuals in a Random Sample of an Animal Population. *Journal of Animal Ecology*,

575   12, 42–58.

576   Frank, S.A. (2009). The common patterns of nature. *Journal of Evolutionary Biology*, 22, 1563–1585.

577   Frank, S.A. (2019). The common patterns of abundance: the log series and Zipf's law. *F1000Res*, 8, 334.

578   Gaston, Kevin J, Blackburn, Tim M, & Lawton, John H. (1993). Comparing Animals and Automobiles: A

579   Vehicle for Understanding Body Size and Abundance Relationships in Species Assemblages?

580   *Oikos*, 66, 172–179.

581   Gaston, Kevin J & Blackburn, Tim M. (2000). *Pattern and Process in Macroecology*. Blackwell Science Ltd.

582   Gotelli, N.J. & Colwell, R.K. (2011). Estimating species richness. In: *Biological Diversity: Frontiers in*

583   *Measurement and Assessment* (eds. Magurran, A.E. & McGill, B.J.). Oxford University Press,

584   Oxford, United Kingdom, pp. 39–54.

585   Haegeman, B. & Loreau, M. (2008). Limitations of entropy maximization in ecology. *Oikos*, 117, 1700–

586   1710.

587   Harte, J. (2011). *Maximum Entropy and Ecology: A Theory of Abundance, Distribution, and Energetics*.

588   Oxford University Press.

589   Harte, J. & Newman, E.A. (2014). Maximum information entropy: a foundation for ecological theory.

590   *Trends in Ecology & Evolution*, 29, 384–389.

591   Harte, J., Zillio, T., Conlisk, E. & Smith, A.B. (2008). Maximum Entropy and the State-Variable Approach to

592   Macroecology. *Ecology*, 89, 2700–2711.

593   Jaynes, E.T. (1957). Information Theory and Statistical Mechanics. *Phys. Rev.*, 106, 620–630.

594 Keil, P., MacDonald, A. a. M., Ramirez, K.S., Bennett, J.M., García-Peña, G.E., Yguel, B., *et al.* (2018).

595   Macroecological and macroevolutionary patterns emerge in the universe of GNU/Linux operating

596   systems. *Ecography*, 41, 1788–1800.

597 Lawton, J.H. (1999). Are There General Laws in Ecology? *Oikos*, 84, 177.

598 Levin, S.A. (1992). The Problem of Pattern and Scale in Ecology: The Robert H. MacArthur Award Lecture.

599   *Ecology*, 73, 1943–1967.

600 Locey, K.J. & White, E.P. (2013). How species richness and total abundance constrain the distribution of

601   abundance. *Ecology Letters*, 16, 1177–1185.

602 Magurran, A.E. & Henderson, P.A. (2003). Explaining the excess of rare species in natural species

603   abundance distributions. *Nature*, 422, 714–716.

604 Maurer, B.A. (1999). *Untangling ecological complexity : the macroscopic perspective*. University of

605   Chicago Press.

606 Maurer, B.A. & McGill, B.J. (2011). Measurement of species diversity. In: *Biological Diversity: Frontiers in*

607   *Measurement and Assessment* (eds. Magurran, A.E. & McGill, B.J.). Oxford University Press,

608   Oxford, United Kingdom, pp. 55–61.

609 McGill, B. (2003). Strong and weak tests of macroecological theory. *Oikos*, 102, 679–685.

610 McGill, B.J. (2019). The what, how and why of doing macroecology. *Global Ecology and Biogeography*, 28,

611   6–17.

612 McGill, B.J., Etienne, R.S., Gray, J.S., Alonso, D., Anderson, M.J., Benecha, H.K., *et al.* (2007). Species

613   abundance distributions: moving beyond single prediction theories to integration within an

614   ecological framework. *Ecology Letters*, 10, 995–1015.

615 Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A. & Leisch, F. (2019). *e1071: Misc Functions of the*

616   *Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*.

617   Nee, S., Harvey, P.H., May, R.M. & Krebs, J.R. (1991). Lifting the veil on abundance patterns. *Proceedings*

618   *of the Royal Society of London. Series B: Biological Sciences*, 243, 161–163.

619   Nekola, J.C. & Brown, J.H. (2007). The wealth of species: ecological communities, complex systems and

620   the legacy of Frank Preston. *Ecology Letters*, 10, 188–196.

621   O'hara, R.B. (2005). Species richness estimators: how many species can dance on the head of a pin?

622   *Journal of Animal Ecology*, 74, 375–386.

623   Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., *et al.* (2020). *vegan:*

624   *Community Ecology Package*.

625   Phillips, O. & Miller, J.S. (2002). *Global patterns of plant diversity: Alwyn H. Gentry's forest transect data*

626   *set*. Missouri Botanical Press.

627   Preston, F.W. (1948). The Commonness, And Rarity, of Species. *Ecology*, 29, 254–283.

628   Preston, F.W. (1950). Gas Laws and Wealth Laws. *The Scientific Monthly*, 71, 309–311.

629   Preston, F.W. (1962a). The Canonical Distribution of Commonness and Rarity: Part I. *Ecology*, 43, 185–

630   215.

631   Preston, F.W. (1962b). The Canonical Distribution of Commonness and Rarity: Part II. *Ecology*, 43, 410–

632   432.

633   Preston, F.W. (1980). Noncanonical Distributions of Commonness and Rarity. *Ecology*, 61, 88–97.

634   Sauer, J.R., Link, W.A., Fallon, J.E., Pardieck, K.L. & Ziolkowski, D.J. (2013). The North American Breeding

635   Bird Survey 1966–2011: Summary Analysis and Species Accounts. *North American Fauna*, 1–32.

636   Shockley, W. (1957). On the Statistics of Individual Variations of Productivity in Research Laboratories.

637   *Proc. IRE*, 45, 279–290.

638   Supp, S.R. & Ernest, S.K.M. (2014). Species-level and community-level responses to disturbance: a cross-

639   community analysis. *Ecology*, 95, 1717–1723.

640   Thibault, K.M., Supp, S.R., Giffin, M., White, E.P. & Ernest, S.K.M. (2011). Species composition and

641   abundance of mammalian communities. *Ecology*, 92, 2316–2316.

642   Ulrich, W., Ollik, M. & Ugland, K.I. (2010). A meta-analysis of species–abundance distributions. *Oikos*, 119,

643   1149–1155.

644   White, E.P., Thibault, K.M. & Xiao, X. (2012). Characterizing species abundance distributions across taxa

645   and ecosystems using a simple maximum entropy model. *Ecology*, 93, 1772–1778.

646   White, E.P., Thibault, K.M. & Xiao, X. (2016). Data from: "Characterizing species abundance distributions

647   across taxa and ecosystems using a simple maximum entropy model". Figshare. Available at:

648   https://doi.org/10.6084/m9.figshare.c.3304845.v1.

649   Woudenberg, S.W., Conkling, B.L., O'Connell, B.M., LaPoint, E.B., Turner, J.A. & Waddell, K.L. (2010). The

650   Forest Inventory and Analysis Database: Database description and users manual version 4.0 for

651   Phase 2. *Gen. Tech. Rep. RMRS-GTR-245. Fort Collins, CO: U.S. Department of Agriculture, Forest*

652   *Service, Rocky Mountain Research Station. 336 p.*, 245.

653   Xiao, X., O'Dwyer, J.P. & White, E.P. (2016). Comparing process-based and constraint-based approaches

654   for modeling macroecological patterns. *Ecology*, 97, 1228–1238.

655   Yenni, G., Adler, P.B. & Ernest, S.K.M. (2012). Strong self-limitation promotes the persistence of rare

656   species. *Ecology*, 93, 456–461.

657

658

659

660   **Figure legends**

661   Figure 1. Distribution of communities from each dataset in terms of total abundance (N) and species

662   richness (S). Communities range from few species and individuals (lower left corner) to speciose

663   communities with many individuals (upper right). However, datasets are not evenly distributed across this

664   state space due to differences in their sampling intensity, design, and underlying biology

665   (e.g. productivity, regional richness, taxonomic group, or other factors that influence the capacity of a

666   community to support large numbers of species and individuals). In particular, note that the FIA dataset

667   comprises very small communities, and communities from the Gentry dataset are extreme in both their

668   high species richness and low average abundance.

669   Figure 2. Large feasible sets may allow better detection of deviations from the statistical baseline by

670   generating more specific, narrowly-defined baselines. We illustrate this phenomenon using 3 hypothetical

671   communities: a small community ($S = 4$, $N = 34$; top row), an intermediate community ($S = 13$, $N = 315$;

672   middle row), and a large community ($S = 44$, $N = 13360$; bottom row). The large communiity has

673   approximately 6.59e+70 possible SADs in its feasible set, while the intermediate community has

674   1.001e+12 and the small community has only 297. For every SAD sampled from the feasible set (left

675   column), we calculate the skewness (color scale) or other summary metrics (not shown). The distributions

676   of these values (right column) constitute the statistical baseline. We define a "breadth index" as the ratio

677   of the range encompassed in the two-tailed 95% density interval (distance between red lines, right),

678   compared to the full range of values for the statistic (distance between the maximum and minimum

679   values). As S and N increase, the size of the feasible set increases, resulting in a narrower statistical

680   baseline (smaller breadth index) – thus enabling easier detection of deviations that may be the result of

681   ecological processes affecting the SAD.

682

683   Figure 3. Many ecological communities are highly unusual compared to their statistical baselines.

684   Percentile ranks are calculated by comparing each community to its sampled feasible set, with very high

685   or very low percentile ranks reflecting extreme values relative to statistical baselines. The vertical red

686   lines mark the 95th percentile for the dissimilarity to the central tendency, and the 2.5th and 97.5th

687   percentiles for all other metrics. Species abundance distributions that are sampled at random from the

688   feasible set will produce percentile ranks that are roughly uniformly distributed from 0 to 100, with

689   approximately 5% of values above the 95th percentile or 2.5% of values above and below the 2.5th and

690   97.5th percentiles, respectively. In contrast, most datasets have more communities that are highly skewed

691   or uneven than would be expected by chance. The percentile values shown are the mean of the

692   percentile scores defined as the proportion of comparison values <=, and <, the focal value. In calculating

693   the proportion of sites with extreme values, the <= designation gives an appropriately conservative

694   estimate of the proportion of high values, but overestimates the proportion of very low values, and the

695   reverse occurs for the < designation.

696   Figure 4. The variability of the feasible set, defined as either the 95th percentile of dissimilarity scores

697   comparing elements of a feasible set to the central tendency of that feasible set, or via a breadth index

698   (see Figure 2), decreases as the number of possible SADs in the feasible set becomes very large. Highly

699   variable feasible sets constitute broad, poorly-defined statistical baselines that may impede our ability to

700   confidently detect deviations between observations and what is expected given the baseline. Small

701   feasible sets, which occur for small combinations of S and N, are often highly variable. The majority of

702   these small, highly variable feasible sets occur for communities in the FIA and Mammal Community

703   databases. Although the Gentry dataset also contains communities with small feasible sets, these

704   communities also have a very low ratio of N to S, meaning their entire feasible sets may be constrained to

705   be more self-similar than small feasible sets in general (see Dissimilarity to central tendency). There is,

706   however, substantial additional variation in the dissimilarity and breadth indices not accounted for by the

707   size of the feasible set or the ratio of N to S.

708   Figure 5. Summaries of how resampling to adjust for under-detection of rare species (orange) and

709   subsampling (pink) change the proportion of extreme values observed for each metric and dataset. The

710   horizontal black lines mark the approximate proportions of extreme values that would be expected at

711   random: 5% for dissimilarity to the central tendency, and 2.5% for all other metrics. Adjusting for rare

712   species consistently increases the proportion of extreme values relative to the raw SADs, while

713   subsampling often decreases it but generally does not eliminate or change the direction of the effect. The

714   exception is for the FIA dataset, which does not show strong deviations for either raw or resampled SADs.

715   Shown are the effects and directions observed for most datasets; for complete results of resampling,

716   including the opposite direction effects, see A7.

717   Figure 6. Very small communities (e.g. those with fewer than 2000 possible SADs in the feasible set;

718   upper rows) exhibit more variable, broadly-defined statistical baselines (top) and less consistently

719   extreme observed values relative to their feasible sets (bottom). 2000 possible SADs is used as a cutoff

720   because it allows for a comparison using a substantial number of communities from the FIA and two

721   other datasets. Of these datasets, the FIA is the most dominated by very small communities (68% of FIA

722   sites have fewer than 2000 possible SADs, compared to 34% for the Mammal Community and 7% for the

723   Miscellaneous Abundance databases). Results shown are for skewness; for complete results see Appendix

724   A10.

725

726