Ecology Letters, (2003) **6**: 766–773 doi: 10.1046/j.1461-0248.2003.00491.x

REPORT

Does Mother Nature really prefer rare species or are log-left-skewed SADs a sampling artefact?

Brian J. McGill
Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721, USA
Correspondence: E-mail: mail@brianmcgill.org

Abstract

Intensively sampled species abundance distributions (SADs) show left-skew on a log scale. That is, there are too many rare species to fit a lognormal distribution. I propose that this log-left-skew might be a sampling artefact. Monte Carlo simulations show that taking progressively larger samples from a log-unskewed distribution (such as the lognormal) causes log-skew to decrease asymptotically (move towards $-\infty$) until it reaches the level of the underlying distribution (zero in this case). In contrast, accumulating certain types of repeated small samples results in a log-skew that becomes progressively more log-left-skewed to a level well beyond the underlying distribution. These repeated samples correspond to samples from the same site over many years or from many sites in 1 year. Data from empirical datasets show that log-skew generally goes from positive (right-skewed) to negative (left-skewed) as the number of temporally or spatially replicated samples increases. This suggests caution when interpreting log-left-skew as a pattern that needs biological interpretation.

Keywords

Left-skew, species abundance distributions, sampling.

Ecology Letters (2003) 6: 766–773

INTRODUCTION

Most species are scarce. Plotting a histogram of the abundances of different species within a community makes this obvious. Scientists call this plot a species abundance distribution (SAD). SADs invariably display a strongly right-skewed pattern known as a hollow curve (Fisher *et al.* 1943; Preston 1948, 1962; Whittaker 1965; May 1975; Brown 1995; Gaston & Blackburn 2000). A SAD describes compactly the structure of a community, so understanding the causes of SADs may tell ecologists a great deal about how communities are structured.

In 1948, Preston proposed plotting a histogram of log-transformed abundances instead of arithmetic abundances. He discovered that the pattern on a log-scale is modal (humped) and appears similar to a normal or Gaussian distribution (e.g. Fig. 1). This would make the SAD lognormal. But, because we do not observe very rare species, the left end of the distribution appears chopped off or truncated. Preston called this the 'veil-line'. In the 1960s MacArthur (Hutchinson 1967, p. 362), based on a hint of a pattern in empirical data, suggested that if we could lift the veil we would see a log-left-skewed distribution. Skew measures asymmetry and one calculates skew as the third central moment divided by the third power of the standard deviation. A left-skewed distribution has negative skew, i.e. a long and/or heavy left tail (relative to the right tail), and the mean occurs to the left of (smaller) than the median and the mode (e.g. Fig. 1). A log-left-skewed distribution has negative (left) skew on a logarithmic scale (but may in fact have right skew on an arithmetic scale, as do most SADs).

Considerable empirical work on a diverse group of organisms shows both Preston and MacArthur right. Nearly all sampled communities of more than a few species demonstrate a modal histogram on a log-scale appearing nearly lognormal. Studies have shown that increasing sampling intensity lifts the veil (moves it to the left) and we observe progressively rarer species. Recent work on intensively sampled data shows SADs often are log-left-skewed (Nee *et al.* 1991; Gregory 1994, 2000; Gaston & Blackburn 2000; Hubbell 2001; Magurran & Henderson 2003). These studies show considerable variation in the log-skew, with some showing right log-skew and many showing left log-skew that is not statistically significant. But in the end, intensively sampled

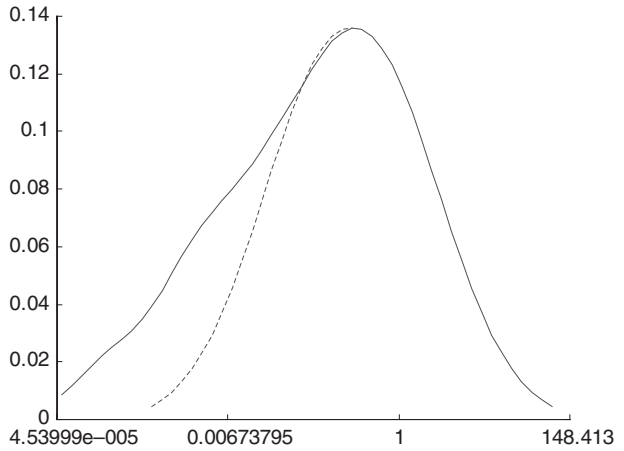

Figure 1 Example of log-left-skew. This figure shows a SAD for bird abundances in the North American BBS study area (see methods). I used kernel smoothing on the SAD to to eliminate noise. I averaged this sample over 5 years and all 1401 good sites. The dotted line is a reflection of the right side of the distribution. Thus, the left tail is clearly higher than the right tail. In fact 12.7% of the probability density lies in the region between the reflection (dotted) line and the observed (solid line) left tail. Scientists usually consider log-left-skew (i.e. an enlarged left tail) as a sign that rare species are more common in nature (even on a log abundance scale).

SADs clearly show a pattern of log-left-skewness on average which often proves statistically significant. Many scientists now consider log-left-skew a benchmark of theories purporting to explain SADs; theories must now reproduce this left skew on log scale to be taken seriously (Nee *et al.* 1991; Harte *et al.* 1999; Hubbell 2001). Presumably, these scientists believe that this log-left-skew represents the underlying natural world – that Mother Nature possesses a bias toward rare species. Many ecologists also feel relief at rejecting a simple lognormal distribution (which has no skew), since the lognormal distribution does not require a biological mechanism to explain the SAD pattern (May 1975; McGill 2003) despite the SADs fundamental importance.

I propose that the log-left-skew commonly found in intensively sampled SADs need not signal the true underlying distribution, but could instead occur as a simple artefact of the intensive sampling needed to lift the veil. I demonstrate the potential for this effect by two simple Monte Carlo simulations. The first looks purely at unveiling by comparing progressively larger samples. The second simulation looks at the effect of accumulating many small samples. Using empirical data, I show that skew becomes increasingly more negative (left-skewed) as samples are added across space or time in a fashion similar to the second model.

METHODS AND MATERIALS

Empirical data

Two empirical datasets were used to test the applicability of the models. The first data set is the North American Breeding Bird Survey (BBS) (Robbins *et al.* 1986; Sauer *et al.* 1997). This survey takes counts of all birds seen or heard at 50 3-min stops along a 24.5 mi (40 km) route. Volunteers survey thousands of routes during the breeding season in the continental US and southern Canada. This process is repeated every year, with the history of some routes going back over thirty years. I used 1401 routes that were rated as 'good quality' by the administrators based on criteria such as time of day and weather conditions for all 5 years from 1996 to 2000. Except where I report data by year, I use data averaged across these same 5 years (1996–2000) to eliminate noise.

The second empirical dataset is a census of tropical trees found on a 50 ha plot on Barro Colorado Island (BCI) in Panama (for detailed methods see Pyke *et al.* 2001; Condit *et al.* 2002). This data was taken during a single year. Data is available by individual 1 ha subplots.

Regional pool

For both Monte Carlo models I assume a two level structure. The higher level is the regional pool (Ricklefs 1987) or metacommunity (sensu Hubbell 2001). The lower level is called the local community. A single census measures a local community. The regional community cannot be directly measured.

Both Monte Carlo models share a single regional pool. Before generating the regional pool, I set the species diversity, called *S*. The regional pool is then generated by drawing *S* abundances (and hence *S* species) with 15 digits of precision *S* times from a lognormal distribution. Note that the lognormal distribution is a continuous distribution and thus I am effectively assuming an infinite number of individuals. On average, this process produces an unskewed SAD. In all cases, I chose the parameters of the lognormal distribution (Evans *et al.* 1993) to be $\mu = 15.45$ and $\sigma = 1.30$, so as to match the estimated lognormal parameters for the total abundance of birds in the BBS dataset. In both Monte Carlo models, I then select a series of samples from this regional pool, drawing whole (discrete) individuals at random in each sample (with replacement). Sensitivity analysis performed using either a logseries distribution for the regional pool (and discrete individuals) or sampling without replacement show no difference in the qualitative outcome (approximate degree of log-left-skew generated). Because the lognormal distribution has no skew, I use the lognormal throughout the rest of the paper. I also used sampling

with replacement in the rest of this paper, because it is computationally faster and better fits the assumption of the regional pool being infinite in size. Sensitivity analysis on the size of the regional pool, S, also showed no effects, and except where stated otherwise, I used $S = 500$.

Model I – unveiling via large sample

In the first simulation, I explore the effect of sample size of the local community. I draw samples where the local community ranges in size from 100 to 10 000 000 individuals. I call this parameter (size of local community which, in this case, equals size of the sample), N_{LOCAL}. The skew of these samples on a log scale were then calculated. Although only one sequence of increasing local community sizes (N_{LOCAL}) is reported, this process was repeated many times and the same results (excepting small random deviations) were obtained.

Model I – accumulating small samples

In the second set of simulations, I explore the effect of accumulating (through summing or averaging) multiple small samples from the regional pool. I sample a local community from the regional pool, where the size of the local community (and sample size) is again called N_{LOCAL}. I then create a second sample of the same size (N_{LOCAL}). However, this second sample is not a *de novo* independent sample of the regional pool. Instead, it is created by removing a a percentage of the individuals in the first local community and replacing only these removed individuals via a sample from the regional pool. I call the percentage of individuals removed and replaced via resampling %REPLACE. Note that %REPLACE is a measure of autocorrelation or (in the opposite direction) independence between samples. When %REPLACE is 0%, the two samples are identical and completely correlated. When %REPLACE is 100%, the two samples are completely independent with no autocorrelation between samples. I then repeat this process of removing %REPLACE of the individuals in the second sample (local community) and replacing them via sampling from the regional pool to generate a third sample. I repeat this process until I have accumulated the specified number of samples, which I call #SAMPLES. I then sum all #SAMPLES samples together to obtain total abundances for each species found in any of the local communities (usually a small subset of the regional pool). I then calculate the skew on a log scale of this SAD. Note that taking an average instead of a sum does not change the shape (in particular the skew) of the distribution.

RESULTS

Model I – unveiling via large sample

As expected, as one increases the sample size (N_{LOCAL}) we see the veil effect (*sensu* Preston 1948): small samples do not reveal the rare species (Fig. 2). As Preston predicted, the abundance of the rarest observed species does decrease as sample sizes increases (Fig. 2). In particular, in every sample one species occurs which has only one individual, so the veil occurs at $N_{\mathrm{VEIL}} = 1$ or if we express N_{VEIL} as a percentage of the total community then $N_{\mathrm{VEIL}} = 1/N_{\mathrm{LOCAL}}$. However, as noted analytically by Pielou and Dewdney (Pielou 1977; Dewdney 1998), the veil does not work as the simple truncation imagined by Preston. Instead, the shape of the observed distribution of abundances in the local community also changes as Pielou and Dewdney predicted. The initial distributions with small sample size are strongly right-skewed (because of the veil on the left tail) with the skew

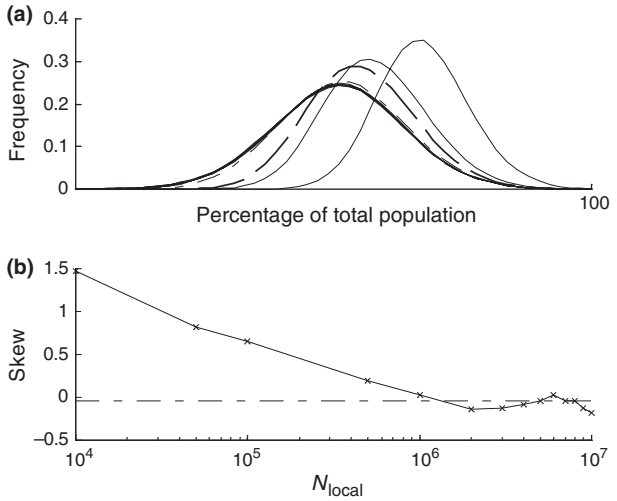

Figure 2 Representative example of the veiling process. (a) A kernel-smoothed frequency distribution for samples ranging in size from 100 to 10 000 000 individuals. The rightmost, most highly-peaked curve is 100. These proceed in order to the leftmost, lowest-peaked curve which is 10 000 000. Notice how the rightmost curve has a heavier right tail, while the leftmost curves have roughly equal tails (zero skew). As Preston noted, there is an unveiling, but as Pielou and Dewdney have demonstrated, the shape of the curve changes as well. In particular, the small sample curves are more right skewed than the underlying curve. (b) A plot of skew vs. sample size. The samples are drawn from a metacommunity which follows a lognormal distribution. Small samples (e.g. 100 individuals) of the metacommunity are highly right skewed (positive). As the sample grows (e.g. 1 000 000) the skew approaches the skew of the metacommunity (horizontal line and here slightly less than 0 by chance). Further sampling causes variation in skew but it remains centred around the skew of the original metacommunity distribution.

dropping down to the skew of the original metacommunity (0 on average for the lognormal) as the sample size approaches the size of the metacommunity (Fig. 2).

Model II – accumulation of small samples

Under the model of accumulation of small samples, skew continues to become more negative. The resultant skew is on average markedly more log-left-skewed than the log-unskewed lognormal distribution and usually falls into the range commonly reported on empirical data (roughly −0.2 to −0.4).

I now report the effect of each of the four parameters in the model (*S*, #SAMPLES, N_{LOCAL} and %REPLACE) on final log-skew. The number of species in the regional pool, *S*, over a range of 50–1000 species had no effect (one-way ANOVA, $P = 0.50$). Repeated model runs show that log-left-skew increased markedly (log-skew became more negative) as #SAMPLES increased from 1 to 10, slowing down as #SAMPLES approached 15, but still decreasing (becoming more log-left-skewed) somewhat out to #SAMPLES = 30. Therefore, I report all further results in this paper using #SAMPLES = 15. Beyond this effect of needing a certain minimum number of repeated samples, #SAMPLES had little effect.

The parameters that had the largest effect were N_{LOCAL} and %REPLACE. To explore the effect of these parameters, I fixed *S* at 500 and #SAMPLES at 15, and I then ran 100 Monte Carlo simulations with a different random seed for each combination of the two remaining parameters (N_{LOCAL} = 100, 500, 1000, 5000, 10 000, 100 000 and %REPLACE = 0.01, 0.05, 0.1, 0.15, 0.2, 0.3, 0.6, 1.0) in a fashion similar to a two-way ANOVA with 100 replicates. The effects of both parameters show high statistical significance (two-way ANOVA, both $P < 0.001$) as does the interaction term ($P < 0.001$). Skew as a function of either parameter shows a 'check-mark pattern' (Fig. 3), where log-skew becomes more negative (log-left-skewed) as either parameter increases initially, reaches a minimum (lowest log-skew) at still fairly small values of the parameter, and then increases (becomes less log-left-skewed, approaching zero-log-skew) over the rest of the parameter range. I found the largest absolute log-skew (most log-left-skewed) for intermediate values of N_{LOCAL} (around 500–1000) and for %REPLACE around 5–10% (Fig. 3). When %REPLACE equals 100%, the final log-skew is very small (absolute value usually <0.02), and is statistically indistinguishable from 0, the log-skew of the regional pool. This makes sense, as the case where %REPLACE = 100% equates mathematically to the unveiling scenario described previously – it is simply the accumulation of independent individuals with no autocorrelation between samples. If the replacement individuals come from within the surviving local community rather than from the regional pool (e.g. via births), the degree of log-left-skew increases even more (becomes more negative).

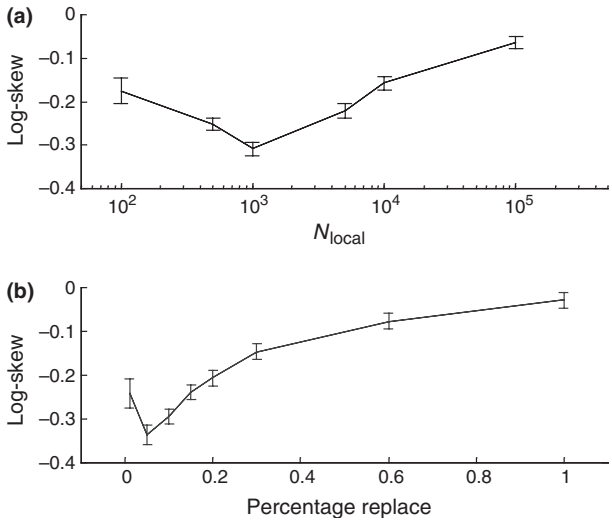

Figure 3 Effect of parameters on degree of log-left-skew. (a) shows the effect of N_{LOCAL} (note the log-scale on the *x*-axis) and (b) shows the effect of %REPLACE. Both demonstrate a 'checkmark' pattern. Log-skew is most negative (made most left-skewed) for intermediate parameters. The effects of both parameters are significant ($P < 0.001$). For the most part the interaction of the two terms is additive, except that when N_{LOCAL} is very small ($N_{\mathrm{LOCAL}} = 100$) and and %REPLACE is very small (%REPLACE = 0.01, 0.05), then the interaction term is negative ($P < 0.001$) giving even lower skew values. For these figures, $S = 500$, SAMPLES = 15, and the remaining parameter [%REPLACE in (a), N_{LOCAL} in (b)] varies across a wide range of values.

In general, when %REPLACE falls in the range 5–20%, N_{LOCAL} falls in the range 500–10 000, and #SAMPLES exceeds 10, the final log-skews are in the range of −0.20 to −0.50, despite their being drawn from a regional pool with no log-skew (on average). The range −0.20 to −0.50 corresponds well with the range of negative skews found in empirical datasets (Nee *et al.* 1991; Gregory 1994, 2000).

Empirical data

Both data on birds from the BBS and data on trees from the BCI 50-ha plot demonstrate the effect of spatial and temporal autocorrelation and repeated sampling (Fig. 4). The average route in the BBS data has an arithmetic skew of mean ± SD = 3.18 ± 1.14 and a log-skew of 0.11 ± 0.26 (i.e. log-right-skewed). The log-skew of the sample pooled across all sites is −0.349 (i.e. log-left-skewed), quite similar to the values found by Gregory for birds in Europe (2000). Thus, on average, each route starts with a strong right skew, with some right skew appearing even on a log scale. Despite

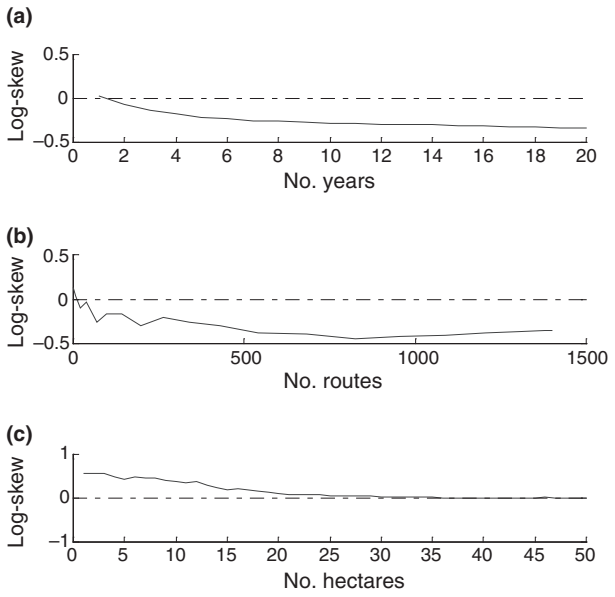

Figure 4 Changes in species abundance distribution (SAD) skew as samples are repeated over time or space. (a) Looking at abundance across all routes studied in the BBS data, the skew goes from slightly positive to strongly negative as we go from sampling one to sampling 20 years. (b) Again in the BBS data, this same pattern occurs (with more noise) as routes (space) are added. In both cases, the initial skew is slightly positive and ends up with a final skew of *c.* −0.35. The exact nature of this graph depends on which route is used to start with. Certain routes along the coast display a more complicated pattern. (c) A similar graph for the BCI data. As plots are added (each 1 hectare in size), the skew goes from strongly positive (log-right-skewed) to very slightly log-left-skewed (−0.0029). It is unclear whether the BCI data demonstrates the unveiling scenario or the repeated sampling scenario as described in this paper.

this, log-skew decreases (becomes more left-log-skewed) as the number of spatially or temporally replicated samples increases (see Fig. 4). Similar results are observed for increases in spatial scale for the BCI data (see Fig. 4), although in this case log-skew merely approaches zero (no log-skew, rather than a negative or left log-skew).

DISCUSSION AND CONCLUSIONS

Causes of log-skew in models I and II

The cause of log-skew dropping towards zero (unskewed) in the first Monte Carlo simulation (unveiling) is fairly obvious: we find rare species only in large samples, truncating the left tail and making the right tail larger in comparison. Increasing sample size merely removes the log-right-skew originally caused by small sample size.

The cause of the increasingly more negative (left) log-skew in the second Monte Carlo simulation (accumulation of small samples) is less obvious. The mechanism for the repeated samples case depends on autocorrelation between the repeated samples (as evidenced by the disappearance of log-left-skew below zero when %REPLACE increased to 100% and by the increase in log-left-skew when replacements came from within the remaining local community instead of the regional pool). The regionally common species that make it into the local pool usually remain in the local pool through all the replicated samples because of their commonness within the local community and the autocorrelation. Those rare species that do make it into the local community very often disappear after just a few repeated samples because of their rarity (i.e. they are more likely to have all individuals replaced than a common species). See Fig. 5. The relative abundance of these species (as a proportion of the total community) then continues to decrease as the number of samples increases, only because of their absence from the additional samples. This causes the rare species to have extremely low abundances. This creates the log-left-skew in excess of that found in the regional pool.

This same type of autocorrelation has been described in a different context as the 'positive occupancy–abundance

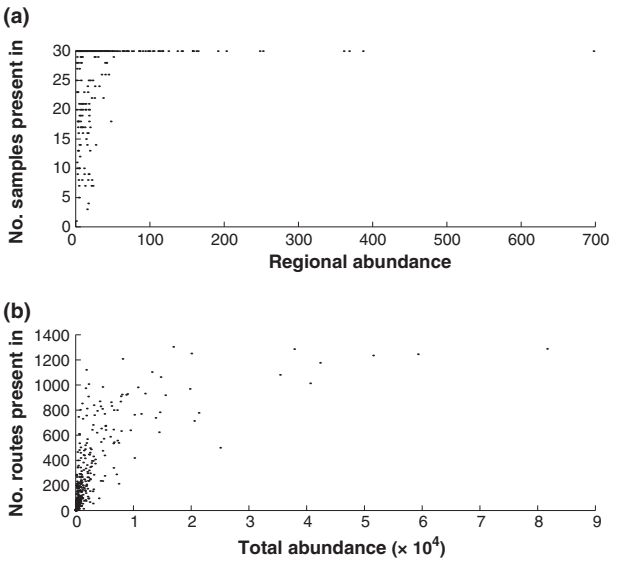

Figure 5 Interaction between regional abundance and the number of samples in which a species is found. Very abundant species are found in every sample (all 15 samples). Rarer species are much more likely to be found in only a few samples. Species found in zero samples are omitted (but are found at all levels of regional abundance). The top figure (a) is from a simulation run of the model II (accumulation of small samples). The bottom figure (b) is for BBS data. When abundance is transformed to a log scale, these graphs produce data that appears similar to a logistic curve. The pattern shown in these graphs is closely related to the positive occupancy-abundance relationship as mentioned in the text.

relationship' (Hanski 1982; Brown 1984, 1995; Gaston *et al.* 1997, 2000).

Relevance of models to empirical data

As this repeated sampling effect can generate strongly log-left-skewed populations even when sampling from unskewed populations, one must ask whether the mechanism is biologically relevant. The fact that only very intensively sampled datasets show log-left-skew suggests that model II (accumulation of small samples) might be relevant. The sampling intensity is generated by repeated samples in space or in time, or often both space and time. For example, data on birds at the scale of a nation make up the majority of claims for left skew (Nee *et al.* 1991; Gregory 1994; Gregory & Gaston 2000). This data always derives from surveys conducted across many years and at many sites.

The second model (accumulated samples) was specifically designed to simulate the biological process involved when samples are accumulated from a given spatial location (local community) over time. Individuals die (%REPLACE) and are replaced. In the case where replacements come from the regional pool, this represents migration. In the case where the replacements come from the remaining local pool, this represents births. #SAMPLES in this scenario equates to the number of years.

Moreover, the empirical data from the BBS confirms that log-skew becomes increasingly negative (log-left-skewed) as the number of years increases in a fashion very similar to that in which the log-skew becomes more negative in the simulations as #SAMPLES increases.

The relevance of the second model is more subtle in cases where sampling intensity is achieved by increasing spatial scale instead of time. Individuals do not remain to be sampled in ensuing samples as one moves across space. However, there is a well-documented pattern of gradual decay of similarity of species abundance and composition over space (Rosenzweig 1995; Condit *et al.* 2002) which creates an autocorrelation between samples very similar to that generated by my model II (accumulated small samples). In particular, common species disappear from ensuing samples much less frequently than rare spaces (in Fig. 5 compare the top figure, i.e. from the model, and the bottom figure, i.e. from empirical data accumulated across space). Thus, any process which creates partial autocorrelation between samples with rare species more likely to disappear from ensuing samples will generate this log-left-skew as repeated sampling occurs.

Related work

Gotelli & Colwell (2001) have pointed out that sampling can produce many pitfalls in the estimation of species richness as well. Several authors have pointed out that the shape of the SAD changes with scale. Whittaker (1965) suggested that the shape of the SAD depends on taxonomic scale (number of species). Tokeshi (1993) cautioned that the shape of the SAD might depend on spatial and temporal scale, and Wilson and coworkers (Wilson *et al.* 1996, 1998) confirmed this empirically. Other authors have pointed out that a variety of patterns change with scale and have corresponding changes in the causal processes with scale (Levin 1992; Rosenzweig 1995). However, I suggest that this work should not be confounded with the important issues of scale. For example, Fig. 4 shows that there are marked changes in log-left-skew simply in going from 2 to 5 years or 5 to 15 ha. Most people would not consider this a change in scale, merely an incremental increase in time and space. This model does not address the issue of whether the shape of the SAD might vary with scale. This model does suggest that even at a single scale, the autocorrelated nature of accumulated, non-independent samples can artificially create the appearance of changing shape of the SAD.

This work shares some similarities to that of Hubbell (2001), starting with the idea of a two-level community structure (local and regional). Hubbell's model also emphasizes dynamic turnover (replacement of individuals) in the local community as a central process. However, Hubbell emphasizes the continuously varying parameter, m, which measures the amount of replacement from the local community vs. the regional pool, while I only modelled the two extremes ($m = 0$, and 1). In contrast, I emphasize the continuously varying parameter for rate of turnover (%REPLACE), while Hubbell usually de-emphasizes this parameter (one turnover per time step). Despite these differences in emphasis, it is interesting to note that Hubbell also points out that log-skew changes as a result of variation in this process of local turnover and replacement (Hubbell 2001, p. 133). However, he has rare species disappearing in the underlying distribution through dispersal limitation (parameter m), while I have it occurring because of autocorrelation between samples. These two effects may ultimately trace back to the same underlying biology. It should be noted that although Hubbell (2001) cites the BCI data as an example of log-left-skew (p. 134), it in fact has a log-skew quite close to zero (−0.0029). Note that because I used a dataset from the same location but different years, the quantitative results will vary slightly but the shape of the SAD does not change noticeably. This log-skew of zero is because the distribution is truncated with a minimum abundance of one individual. If one imagines the left-tail being extended out, then the SAD does appear as if it would be left-skewed (Hubbell's Fig. 5.7).

Several authors working with large empirical datasets have arrived at the conclusion that accidental species cause the log-left-skew. Gregory (2000) noted that the removal of

10 accidental species (predominantly African or Asian in origin) eliminated the large log-left-skew observed in Europe as a whole (i.e. made log-skew equal to zero). Accidental species are species that are not native to the area under observation and are observed only occasionally (i.e. in a few years). Gregory's empirical observation supports the simulation results in this paper where log-left-skew is created by species that appear in only a few of the repeated samples. Magurran & Henderson (2003) found a very similar result in a 21-year dataset on an estuarine fish community. They found that 'core' species that are 'persistent, abundant and biologically associated with estuarine habitats' follow a lognormal distribution (with zero log-skew), while the 'occasional' species which occur infrequently in the record follow a logseries distribution which is strongly left-skewed on a log-scale. Adding these two groups together produces a log-left-skewed distribution. This again supports the results given by model II (accumulation of small samples) where log-left-skew is generated by species that occur in only a few of the samples. Moreover, Magurran and Henderson display a graph which demonstrates an autocorrelation signature very similar to that found in the second model and the BBS data (my Fig. 5 vs. their Fig. 1a if the axes are swapped). Finally, a close analysis of the BCI dataset suggests that the appearance of left-skew is almost entirely the result of species that were observed to have only one individual in the 50 ha plot (and that species with one individual are more common than species with 2, 3 or 4 individuals). Removal of species with only one individual changes log skew from −0.0029 to 0.166. These might well be regarded as accidental or occasional species as well, although a more careful analysis of the biology and the natural habitat affinities of these rare species is necessary to justify calling them accidental. Such careful studies are vitally important, as arbitrarily removing rare species will always make a distribution less left-skewed. Magurran and Henderson give a very careful analysis of the biology that justifies their separation of species as core vs. occasional.

CONCLUSIONS

Four caveats emerge from this work when exploring skew in SADs:

1 Unveiling does not work as a simple gradual revealing of the left side of the distribution. Instead, the whole distribution changes shape as the number of individuals sampled increases. The true nature of unveiling has been noted before (Pielou 1977; Dewdney 1998), but is still often ignored.
2 Purely statistical mechanisms without biology may cause the observed propensity for log-left-skew. Thus, theories do not need to generate log-left-skewed SADs to be considered realistic. Producing lognormal curves for single sample SADs remain consistent with empirical data. This statement should be qualified by recognizing that the causes of spatial and temporal autocorrelation are clearly biological. Another qualification is that there may still be other causes of log-left-skew that indicate that Mother Nature does indeed favour rarer species. This model does not rule them out, it merely argues that we need to use very careful statistics and address the problem of autocorrelated accumulated samples before making such a claim.
3 This paper should serve as a caution to those who wish to devote great effort to distinguishing curves (and associated theories) based on small differences in the tails of a probability distribution. By definition, tails occur where events are rarest and hence most prone to sampling effects. Similarly, although the log transform is a natural transformation for population abundances, it has the effect of emphasizing the left tail at the expense of the rest of the distribution. And despite what I perceive as an emerging 'tyranny of the log scale' for SADs, the right tail drives many important ecological questions (e.g. flow of energy or nutrients through an ecosystem).
4 Empiricists must remember that sampling does skew the underlying distribution: small samples cause artificial right skew and, paradoxically, the more we sample (when in a repeated, autocorrelated fashion) the more we cause log-left-skew distortion and lose information about the underlying distribution. The solution is to follow as much as possible the approach suggested by the first (unveiling) model by taking a single large sample. This approach gives a more accurate view of the underlying distribution than accumulating repeated small samples.

In conclusion, this paper suggests a simple, statistical, non-biological mechanism for the often observed log-left-skew in SADs. This mechanism uses the idea of repeated, auto-correlated sampling. Monte Carlo simulations show that repeated, auto-correlated sampling as described above clearly generates log-left-skew not found in the underlying distribution and hence an excess of rare species (on a log scale). Analysis of empirical data shows that log-skew changes from positive to negative as spatial and temporal replicate samples are added. We know that spatial and temporal autocorrelation occurs in these samples. Thus, the empirical data seem to support the idea that the repeated, autocorrelated sample mechanism of the Monte Carlo simulations could easily cause the log-left-skew found in empirical data. We need to do additional work to rule out this statistical, abiological theory before we claim that Mother Nature prefers rare species.

ACKNOWLEDGEMENTS

I would like to thank Mike Rosenzweig, Will Turner and Jim Brown for discussions that contributed to my understanding of this issue. I would also like to thank Mike, Will, Sarah Marx, three anonymous reviewers and Brian Maurer for reading earlier drafts of this manuscript.

REFERENCES

Brown, J.H. (1984). On the relationship between abundance and distribution of species. *Am. Nat.*, 124, 255–279.

Brown, J.H. (1995). *Macroecology*. University of Chicago Press, Chicago, IL.

Condit, R., Pitman, N., Leigh, E.G., Chave, J., Terborgh, J., Foster, R.B. *et al.* (2002). Beta-diversity in tropical forest trees. *Science*, 295, 666–669.

Dewdney, A.K. (1998). A general theory of the sampling process with applications to the "veil line". *Theor. Popul. Biol.*, 54, 294–302.

Evans, M., Hastings, N. & Peacock, B. (1993). *Statistical Distributions*, 2nd edn. John Wiley & Sons, New York.

Fisher, R.A., Corbet, A.S. & Williams, C.B. (1943). The relation between the number of species and the number of individuals in a random sample from an animal population. *J. Anim. Ecol.*, 12, 42–58.

Gaston, K.J. & Blackburn, T.M. (2000). *Pattern and Process in Macroecology*. Blackwell Science, Ltd, Oxford.

Gaston, K.J., Blackburn, T.M. & Lawton, J.H. (1997). Interspecific abundance-range size relationships: an appraisal of mechanisms. *J. Anim. Ecol.*, 66, 579–601.

Gaston, K.J., Blackburn, T.M., Greenwood, J.J.D., Gregory, R.D., Quinn, R.M. & Lawton, J.H. (2000). Abundance-occupancy relationships. *J. Appl. Ecol.*, 37, S39–S59.

Gotelli, N.J. & Colwell, R.K. (2001). Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecol. Lett.*, 4, 379–391.

Gregory, R.D. (1994). Species Abundance Patterns of British Birds. *Proc. R. Soc. Lond. Ser. B Biol. Sci.*, 257, 299–301.

Gregory, R.D. (2000). Abundance patterns of European breeding birds. *Ecography*, 23, 201–208.

Gregory, R.D. & Gaston, K.J. (2000). Explanations of commonness and rarity in British breeding birds: separating resource use and resource availability. *Oikos*, 88, 515–526.

Hanski, I. (1982). Dynamics of regional distribution: the core and satellite species hypothesis. *Oikos*, 38, 210–221.

Harte, J., Kinzig, A.P. & Green, J. (1999). Self-similarity in the distribution and abundance of species. *Science*, 284, 334–336.

Hubbell, S.P. (2001). *A Unified Theory of Biodiversity and Biogeography*. Princeton University Press, Princeton.

Hutchinson, G.E. (1967). *A Treatise on Limnology*. John Wiley & Sons, New York.

Levin, S.A. (1992). The problem of pattern and scale in ecology. *Ecology*, 73, 1943–1967.

McGill, B.J. (2003). Strong and weak tests of macroecological theory. *Oikos* (in press).

Magurran, A.E. & Henderson, P.A. (2003). Explaining the excess of rare species in natural species abundance distributions. *Nature*, 422, 714–716.

May, R.M. (1975). Patterns of species abundance and diversity. In: *Ecology and Evolution of Communities* (eds. Cody, M.L. & Diamond, J.M.). Belknap Press of Harvard University Press, Cambridge, MA, pp. 81–120.

Nee, S., Harvey, P.H. & May, R.M. (1991). Lifting the Veil on Abundance Patterns. *Proc. R. Soc. Lond. Ser. B Biol. Sci.*, 243, 161–163.

Pielou, E.C. (1977). *Mathematical Ecology*. John Wiley & Sons, New York.

Preston, F.W. (1948). The commonness and rarity of species. *Ecology*, 29, 254–283.

Preston, F.W. (1962). The canonical distribution of commonness and rarity: part I. *Ecology*, 43, 185–215.

Pyke, C.R., Condit, R., Aguilar, S. & Lao, S. (2001). Floristic composition across a climatic gradient in a neotropical lowland forest. *J. Vegetat. Sci.*, 12, 553–566.

Ricklefs, R.E. (1987). Community diversity – relative roles of local and regional processes. *Science*, 235, 167–171.

Robbins, C.S., Bystrak, D. & Geissler, P.H. (1986). *The breeding bird survey: its first fifteen years, 1965–1979*. US Department of the Interior, Fish and Wildlife Service, Washington, DC.

Rosenzweig, M.L. (1995). *Species Diversity in Space and Time*. Cambridge University Press, Cambridge.

Sauer, J.R., Hines, J.E., Gough, G., Thomas, I. & Peterjohn, B.G. (1997). The North American Breeding Bird Survey Results and Analysis. http://www.mp2-pwrc.usgs.gov/bbs/

Tokeshi, M. (1993). Species abundance patterns and community structure. *Adv. Ecol. Res.*, 24, 111–186.

Whittaker, R.H. (1965). Dominance and diversity in land plant communities. *Science*, 147, 250–260.

Wilson, J.B., Wells, T.C.E., Trueman, I.C., Jones, G., Atkinson, M.D., Crawley, M.J., Dodd, M.E. & Silvertown, J. (1996). Are there assembly rules for plant species abundance? An investigation in relation to soil resources and successional trends? *J. Ecol.*, 84, 527–538.

Wilson, J.B., Gitay, H., Steel, J.B. & King, W.M. (1998). Relative abundance distributions in plant communities: effects of species richness and of spatial scale. *J. Vegetat. Sci.*, 9, 213–220.

Editor, B. Maurer
Manuscript received 26 March 2003
First decision made 1 May 2003
Manuscript accepted 1 June 2003