FAIR SAMPLING OF TAXONOMIC RICHNESS AND UNBIASED ESTIMATION OF ORIGINATION AND EXTINCTION RATES

JOHN ALROY

Department of Biological Sciences, Faculty of Science, Macquarie University, Sydney, NSW 2109 Australia

ABSTRACT.—Paleobiologists are reaching a consensus that biases in diversity curves, origination rates, and extinction rates need to be removed using statistical estimation methods. Diversity estimates are biased both by methods of counting and by variation in the amount of fossil data. Traditional counts are essentially tallies of age ranges. Because these counts are distorted by interrelated factors such as the Pull of the Recent and the Signor-Lipps effect, counts of taxa actually sampled within intervals should be used instead. Sampling intensity biases can be addressed with randomized subsampling of data records such as individual taxonomic occurrences or entire fossil collections. Fair subsampling would yield taxon counts that track changes in the species pool size, i.e., the diversity of all taxa that could ever be sampled. Most of the literature has overlooked this point, having instead focused on making sample sizes uniform through methods such as rarefaction. These methods flatten the data, undersampling when true diversity is high. A good solution to this problem involves the concept of frequency distribution coverage: a taxon's underlying frequency is said to be "covered" when it is represented by at least one fossil in a data set. A fair subsample, but not a uniform one, can be created by drawing collections until estimated coverage reaches a fixed target (i.e., until a "shareholder quorum" is attained). Origination and extinction rates present other challenges. For many years they were thought of in terms of simple counts or ratios, but they are now treated as exponential decay coefficients of the kind featuring in simple birth-death models. Unfortunately, these instantaneous rates also suffer from counting method biases (e.g., the Pull of the Recent). Such biases can be removed by only examining taxa sampled twice consecutively, three times consecutively, or in the first and third of three intervals but not the second (i.e., two timers, three timers, and part timers). Two similar equations involving these counts can be used. Alternative methods of estimating diversity and turnover through extrapolation share some of the advantages of quorum subsampling and two-timer family equations, but it remains to be shown whether they produce precise and accurate estimates when applied to fossil data.

INTRODUCTION

THE MODERN discipline of analytical paleobiology arguably stems from two main sources: the collective work of George Gaylord Simpson (e.g., 1944, 1949, 1952, 1960) and a collaborative project now known as the Marine Biological Laboratories simulations (Raup et al. 1973). These approaches were very different: Simpson focused on raw data and Raup et al. focused on a computer simulation. However, both research programs involved using quantitative methods to test general theories about evolutionary laws, an example being the equilibrial model of diversification (MacArthur and Wilson 1967; Sepkoski 1978) that was built in to Raup et al.'s simulation. Because there is no space to talk about testing such hypotheses in this chapter, my much more modest goal is to show how the diversity and turnover variables of interest to Simpson and his followers can be estimated with methods stemming from Raup et al.'s work.

Many sources of data are used by ecologists and evolutionists to study taxonomic diversity. The reason I'm a paleobiologist is that fossils make it possible to study diversity (and other things) by generating linked time series instead of isolated numbers. That is, paleobiologists can construct diversity curves that show counts of taxa in successive geological time intervals – motion pictures of evolutionary processes instead of still photographs.

In a perfect world each diversity curve would fully enumerate everything that ever existed (i.e., the richness of the species pool, which along with evenness of abundance distributions is one component of what most ecologists call "diversity"). This kind of

In Quantitative Methods in Paleobiology, pp. 55-80, Paleontological Society Short Course, October 30th, 2010. The Paleontological Society Papers, Volume 16, John Alroy and Gene Hunt (eds.).

an illustration can only be produced faithfully using fossil data. The alternative would be using much less direct sources of information such as phylogenies of extant species and taxon counts representing different communities.

Unfortunately, fossil-based diversity curves are almost always badly distorted because so many taxa have incompletely sampled geographic, environmental, and temporal distributions or are never sampled at all (Newell 1959; Simpson 1960; Raup 1972). Instead, a curve is at best a numerical estimate of the *relative* amount of diversity in each time interval compared to the others.

Taxonomic turnover rates are even more slippery. For one thing, there is more than one kind of "rate": there are simple counts of taxa first or last appearing in a time interval; percentages; and proper instantaneous rates of the kind I will define below. For another, rates are subject to all of the biases affecting diversity curves and more. A trivial example is that unless the scope of the data are global one cannot be sure that first and last appearances document evolutionary origination events and extinctions instead of immigration events and local extirpations.

The purpose of estimation is to get rid of all these biases. Paleontologists have discussed biases for more than a half century (e.g., Newell 1952), but for several decades all they did was list the possible problems while actually using raw data. Most of these biases somehow related to variation through time in the sheer amount of information. Later, simple corrections for such sample size biases were presented (e.g., Raup 1975, 1976) but no-one found them entirely satisfactory. Still later, paleobiologists collectively gave up on the problem and went back to analyzing raw data (e.g., Sepkoski et al. 1981). Starting in the mid-1990s, the community again began to debate whether and how to remove biases, with many researchers now using data that were at least nominally "corrected" in some way.

In the bulk of this paper I will focus on methods that dampen sample size biases through what is called standardized subsampling. However, I will begin with a discussion of what might seem a simple and not very important step in creating a curve: going from the raw data, whatever that might be, to the actual taxon counts. The choice of a counting method turns out to be potentially very important.

DIVERSITY CURVE COUNTING METHODS

The counts making up diversity curves can be drawn from two basic kinds of data. First, traditional databases consisted of simple lists giving taxon names, first appearances, and last appearances (e.g., Sepkoski 1982, 2002). These appearances were based on expert opinions as expressed in such sources as the Treatise on Invertebrate Paleontology, making it difficult to say how they related to finds of actual fossils. Thanks to the nature of the data, most of the early literature obtained counts simply by adding up the age ranges defined by the appearances. Second, most modern literature depends on taxonomic occurrence data, which consist of records showing which taxa are found in which individual fossil samples (e.g., Niklas et al. 1983). Optimally, each sample represents a narrow horizon exposed in a single geographic location.

These types of data seem very different, but of course an age range also can be computed simply by finding the oldest and youngest occurrence of each taxon. The advantage of having occurrence data is that it lets you show which taxa are *not* found in particular intervals falling inside of their age ranges. They can also tell you *how often* taxa are encountered in each time interval, which will turn out to be crucially important later on. And it turns out that most of the counting methods have severe biases that only can be addressed using occurrence data, as discussed in this section.

Fundamental count categories.—All of the common counting methods involve five distinct categories of taxa that can be separated out using either age ranges or occurrences:

(1) Found both before and after a time interval, i.e., ranging through and sampled (N_{r+}).
(2) Ranging completely through an interval but not sampled, i.e., Lazarus taxa (N_{r-}).
(3) Crossing the interval's bottom (lower) boundary and going extinct, i.e., bottom-only boundary crossers (N_b of Foote 1999).
(4) Originating within an interval and crossing its top (upper) boundary, i.e., top-only boundary crossers (N_t of Foote 1999).
(5) Originating and going extinct immediately, i.e., single-interval taxa ($N_1 = N_{FL}$ of Foote 1999).

These counts were illustrated nicely by Foote (2000), but he chose to lump the first two categories (calling them "N_{bt}") because his focus was on traditional compilations that do not record whether taxa are sampled within their age ranges. He also renamed N_b and N_t, calling them N_{bL} and N_{Ft}, but I'll stick with the short versions. A similar illustration is given here (Fig. 1).

It should be noted that single-interval taxa are sometimes called singletons (e.g., Alroy 2000b; Foote 2000). They should not be, because the term has a different, very common meaning in ecology (i.e., taxa represented by one specimen: Preston 1948), and yet another meaning even within paleobiology (taxa found in one collection: Alroy 1996, 2000b). Mea culpa.

Counting methods.—The most common of the three standard counting methods is range through (RT or N_r), i.e., everything ranging anywhere into a bin:

$$N_{r+} + N_{r-} + N_b + N_t + N_1 \quad (1)$$

A related count is boundary crossers (BC), i.e.,

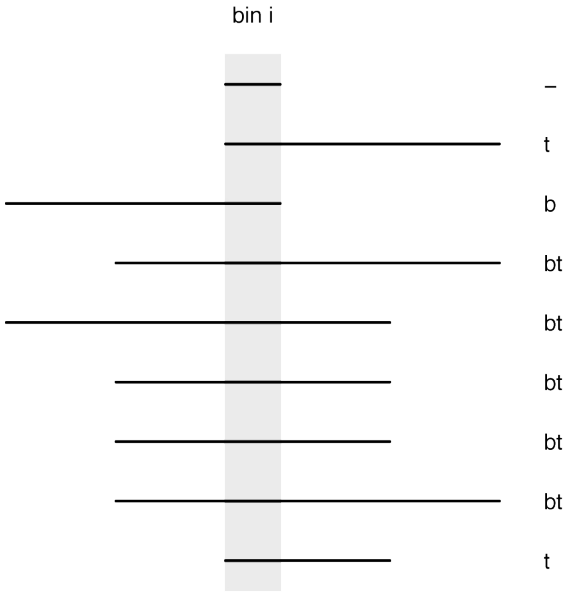

FIGURE *1*.—Examples of age ranges. Each row represents a taxon's known range across five time intervals (bins). Shaded box represents the third bin (*i*). – = restricted to *i*, b = crossing the base of *i* and going extinct, t = originating within *i* and spanning its top, bt = ranging all the way through *i*.

TABLE *1*.—Key to symbols used in the text.

λ	Instantaneous origination rate
μ	Instantaneous extinction rate
N_0	Diversity of a clade at its origination date (= 1)
N_1	One timers (sampled within an interval but not immediately before or after it)
n_1	Singletons (taxa observed once)
N_{2t}	Two timers (sampled immediately before and within an interval)
$N_{2t,i}$	Two timers sampled before and within interval i
$N_{2t,i+1}$	Two timers sampled within and after interval i
N_{3t}	Three timers (sampled immediately before, within, and immediately after an interval)
$N_{3t,i}$	Three timers sampled around interval i
N_b	Bottom-only boundary crossers (sampled anywhere before an interval and last sampled within it)
N_{bt}	Bottom and top crossers (sampled before and after an interval)
N_i	Count of any subset of bottom-only boundary crossers (N_b)
N_{i+1}	Count of any subset of top-only boundary crossers (N_t)
N_{pt}	Part timers (sampled immediately before and after an interval but not within it)
N_r	Range-throughs (ranging anywhere into or across an interval)
N_{r+}	Taxa ranging across an interval and sampled within it
N_{r-}	Taxa ranging across an interval but not sampled within it
N_s	Taxa sampled in an interval (sampled-in-bin diversity)
$N_{s'}$	Standing diversity estimate based on a heuristic rate correction
N_t	Top-only boundary crossers (first sampled within an interval and sampled anywhere after it)
N_R	Diversity of a clade in the Recent
O	Observations (specimens or occurrences)
o_1	Observations of dominant (= most common) taxon
p_1	Taxa only described in a single publication
p_e	Probability of going extinct
p_o	Probability of originating
p_s	Probability of being sampled (computed from N_{3t} and N_{pt})
$p_{survive}$	Probability of surviving into the following interval
q	Shareholder quorum (desired coverage of subsample)
r	Net diversification rate
t	Amount of time between a clade's origin and the Recent
u	Good's estimate of frequency coverage

everything definitely present before and after the base of the bin:

$$N_{r+} + N_{r-} + N_b = N_r - N_t - N_l \quad (2)$$

The most obvious of the three counts is sampled in bin (SIB or N_s), i.e., everything actually sampled:

$$N_{r+} + N_b + N_t + N_l = N_r - N_{r-} \quad (3)$$

Sepkoski (1979, Fig. 6) used the BC measure as part of a calculation, but didn't think it was important on its own. It apparently had not ever been used before, and it was not used to make a diversity curve per se only much later (Alroy 1996; Bambach 1999). Differences between BC and RT have been illustrated by applying them to identical data sets (Bambach 1999; Alroy 2000a). Likewise, BC and SIB were both applied to a Paleobiology Database Phanerozoic invertebrate data set by Alroy et al. (2001).

Many other counts are also mathematically possible but are not widely used. Examples include RT minus single-interval taxa (Sepkoski 1990, 1997); SIB minus single-interval taxa; normalized diversity (RT plus half the rest: Sepkoski 1975, Sepkoski et al. 1981); linear midpoint diversity (the average of BC at the bottom and top: Harper 1975); and "medial" diversity (the mean of bottom BC diversity and RT diversity: Sepkoski 1979, the paper that first mentioned BC in passing). These five methods will not be discussed in more detail because they are either very similar to the RT and BC or, in the case of methods excluding single-interval taxa, can be shown to be severely biased. This is true even though the idea of excluding single-interval taxa was motivated by concerns about sampling bias raised by Foote and Raup (1996) and others.

Biases introduced by counting.—There are five major counting method biases and they have nothing at all to do with sampling biases that are discussed later:

(1) Sampling universe effects. Even if sampling is always fair, sampled diversity will rise whenever the geographic, environmental, or taxonomic sphere of sampling increases. These extra taxa will all be in the single-interval count (N_1) if the sampling universe has been expanded only during one time interval (e.g., because there is a Lagerstätte).

(2) Edge effects (Raup 1972; Foote 2000; Pocock et al. 2004). RT and BC curves drop at their edges because there are many opportunities to sample taxa in the middle, but few at the beginning or end. This matters because we count Lazarus taxa (N_{r-}) as present in a bin even if they are known only from occurrences long before and long after. By definition, this scenario can only occur away from the curve's edges.

(3) The Signor-Lipps effect (Signor and Lipps 1982; Marshall and Ward 1996). Just as with edges, major mass extinctions make it impossible to sample rare taxa that would otherwise be long-ranging. Thus, a curve will drop smoothly as it approaches but does not yet reach an extinction. Likewise, new taxa that appear during the recovery from a mass extinction may take some time to be sampled, so a curve will rise smoothly even if the recovery was extremely fast. The basic problem is that future (or past) sampling events determine whether you infer that the (individually rare) Lazarus taxa are present in any particular bin. They shouldn't.

(4) The Pull of the Recent (Raup 1972, 1978, 1979). The Recent is far better sampled than any interval in the geological record, so ranges of extant taxa are always "pulled forward" to the Recent (unless a taxon is extant but we do not know it yet because it is only known from fossils). Therefore, Lazarus taxa that are lucky enough to be extant are always inferred to be present in a bin if they happen to have a fossil occurrence before it. This bias is the opposite of an edge effect (because it makes the curve go up, not down). The closer you get to the Recent, the more extant taxa there are, so the more important it is.

(5) Rate effects. (a) When rates are high, taxa have short ranges. Short-lived taxa are harder to sample, so RT curves and in particular BC curves are too low when turnover is high. (b) If rates are high more taxa will come and go within particular time intervals (and the same is true if certain intervals are long: Foote 1994). So, total sampled diversity could be considerably higher than diversity at any one point within the same interval (i.e., standing diversity).

By definition, sampled in bin (SIB) diversity cannot have a problem with biases 2, 3, and 4. It is hopelessly challenged by bias 1, especially if the problem is the presence of Lagerstätten. However, that is true of every method, and if the only issue is that some time intervals have been studied more intensely so more is known about their sampling universes, the problem can be eliminated by tweaking sampling methods in a way we will discuss later (Alroy 2010).

Two timers and the rate effect.—At this point we could simply assert that SIB is the only viable option because RT, BC, and all other range-based counts suffer from every single bias and SIB doesn't. However, we have not yet dealt with category 5 (rate effects). We really need to because the average turnover rate has declined greatly through the Phanerozoic marine record (Raup and Sepkoski 1982; Alroy 2008) and because rates are clearly very variable within, say, the Cenozoic record of mammals (Alroy 2000b).

Now, one could show with simple simulations that biases 5a and 5b often cancel out because one makes sampling of standing diversity harder when rates are high and the other makes it easier. That's why SIB has been used without too much concern (Alroy et al. 2008). Ironically, the problem is that it is now possible to produce largely unbiased sampling of species pools regardless of turnover rates, as we will later discuss. When sampling is unbiased SIB is a good proxy for everything-that-ever-existed (= piled up) diversity. If you recall that our goal is to make a motion picture instead of a collage, you can see that's a bad idea. When origination and extinction rates are equal and around 3 taxa per taxon per interval, this pileup can be predicted to be about 30% (Raup 1985, eqn. A29).

It must be stressed that piling up does not occur when turnover is heavily concentrated at boundaries between time intervals (i.e., pulsed). In that case, everything-that-ever-existed is exactly the same as standing diversity. So, pileup is only a big problem when most turnover occurs within intervals, which would be expected if intervals are very long or times are very boring.

There is still not yet a lethal problem because one can often switch to using short intervals, and when times are boring turnover rates are also typically low. Furthermore, some studies suggest that turnover rates actually do accelerate at most boundaries (e.g., Foote 2005; Alroy 2008). In many cases, therefore, we should not expect to see a large rate bias in the first place.

When there is cause to worry about pileup, unfortunately there is no simple, 100% perfect solution to the problem. Any solution claiming to be perfect would probably involve some very nasty likelihood calculations making all sorts of assumptions about processes of turnover and sampling.

Nonetheless, much information about the bias is captured by looking at a single count: the number of taxa sampled immediately before and within a time interval (two timers or N_{2t} : Alroy 2008). N_{2t} is a subset of the number of bottom boundary crossers (Fig. 1), but BC also includes taxa not actually sampled immediately before and after a given bin. If turnover is pulsed, SIB should always equal N_{2t} $(1 + p_o)$ where the count N_{2t} applies to the base of an interval and p_o is the origination probability (as defined below). It also should equal N_{2t} $(1 + p_e)$ where N_{2t} now applies to the top of the interval and p_e is the extinction probability.

Under continuous turnover, these two estimates will differ from each other and from SIB because N_{2t} will be biased downwards, and biased downwards to different degrees in different intervals. This bias occurs exactly because two-timers have to be sampled twice, and when turnover rates are high the short ranges make sampling difficult.

Two-timer counts can still be very informative because they avoid most of the counting problems mentioned above (as does SIB) and are good relative estimates of boundary diversity when turnover is pulsed. However, one would have to be careful about using them because they do compound any sampling error already seen in SIB counts. For example, if you have one really badly sampled bin it will mess up the N_{2t} counts for its own base and for its own top, which is bad enough. But if there are two consecutive bad bins the count for the boundary between them will be multiplicatively bad – potentially zero. The solution to this problem is to make the sampling probabilities uniform (see below).

To sum up, SIB isn't completely reliable, but the traditional range-base counting methods are too biased to be useful. Any improved method would also most likely look at occurrence-based variables such as N_{2t}. Therefore, we should never again use traditional range compilations and should always work with occurrences. Exactly the same conclusion will be drawn from our discussion of sampling intensity biases.

SUBSAMPLING METHODS

Paleontologists have been acutely aware of the fossil record's shortcomings ever since Lyell (1830), and lengthy discussions of this topic have appeared over and over again (e.g., Simpson 1944; Newell 1959; Raup 1972). Sampling biases (broadly including counting method biases) are so widely discussed because the interpretation of diversity curves is a core issue in mac-

roevolution, and arguably the core issue. Researchers taking different positions on whether sampling biases are important also have taken opposed positions on what processes might be governing evolution at the grandest scales (e.g., Valentine 1970 vs. Raup 1972; Sepkoski 1984 vs. Benton 1995).

Early workers such as Simpson (1949) or Newell (1952) understood quite well that even if a diversity curve stems from a comprehensive literature survey that does not mean it is fairly sampled, much less accurate. There was also a vigorous debate over sampling problems in the 1970s. Surprisingly, however, for many decades all parties in such debates did equate having a comprehensive curve with having a good curve – or at least an adequate one. This viewpoint was crystallized in a remarkable paper by Sepkoski et al. (1981) that effectively swept the problem under the rug for well over a decade (Miller 2000).

At this time, however, there is no longer substantive debate about whether comprehensive curves are either fairly sampled or accurate, and if not whether something should be done about it. Instead, a large majority of contemporary large-scale diversity studies employ some kind of algorithm that purports to make sampling fair. These studies employ one of two fundamental strategies. Some of them (e.g., Nichols and Pollock 1983) try to estimate the entire size of the species pool using extrapolation methods, which have been reviewed ably by Colwell and Coddington (1994) and in another chapter in this volume. The others (e.g., Raup 1975; Alroy 1996; Miller and Foote 1996) use standardized subsampling, that is, they count the taxa found in a random subsample of each time interval's data. These studies don't worry about the absolute pool size, but they do try to guarantee that the shape of a curve consistently reflects changes in the pool's size.

My goal is to explain how to get better relative estimates through standardized subsampling without saying very much about extrapolation. I believe there are four really good reasons to subsample instead of extrapolate. First, it's well-understood that most extrapolation methods are noisy, downwards-biased, and dependent on statistical assumptions that are strong and unrealistic (Colwell and Coddington 1994). In other words, they usually don't work. Second, it's also understood that when extrapolation does work, that's because you already have a sample including at least about half the species (Colwell and Coddington 1994). There is usually no reason to think a paleontological data set meets that description. Third, most paleobiologists seem not to care about the absolute pool size. If the goal is test evolutionary theories, it truly is enough to say that (for example) diversity increased or decreased from one time to the next by a certain percentage. Finally, despite making weak assumptions and being very simple, the subsampling method I advocate here demonstrably provides estimates that are both precise and accurate (to the extent allowed by the data). As far as I know, the same has not yet been shown for existing extrapolation methods.

Sources of bias.—Before going into any details we need to clarify what subsampling methods can fix and what they can't. There are many papers giving lengthy laundry lists of biases (e.g., Raup 1972), so I will simply boil them down into two categories.

First, there are biases that have nothing to do with taxonomic attributes and everything to do with context: the original lack sediments or subsequent erosion, metamorphism, and subduction of sediments in certain places; the difficulty of collecting in certain geographic regions; or the lack of interest in certain time intervals. These factors generally have to with how much data and not what kind of data are available, and they are the motivation for trying to impose uniform sampling – however that might be done and regardless of whether uniform sampling is good sampling (I will show it is not).

Second, there are taxon-specific factors that keep things out of the published fossil record, such as the lack of hard parts, the difficulty of identifying what parts do get preserved, the ease of collecting and preparing certain kinds of fossils, or the lack of research interest in certain groups. These problems are devastating for raw enumerations and not easily fixed by subsampling. In fact, I will argue that only one method can do anything about them at all, and that method works best only when we are willing and able to create separate standardized curves for each major taxonomic group (e.g., Alroy 2010).

Accumulation curves and subsampling curves.—Now to methods you might actually use.

Obviously, looking at more fossils generally means finding more kinds of fossils. If that's not obvious, consider that finding a new fossil can add to

the total number of known species but can't subtract from it. However, we will eventually run out of new fossil species to discover as we keep looking at more fossils. In between, discovery of new species will become harder and harder because the common ones will be found first. Therefore, we expect to see a curved, asymptotic relationship between the amount of fossil data and perceived fossil diversity.

A curve showing the accumulation of fossils and fossil species through historical time is called a accumulation curve (Fig. 2.1). A curve showing how ever increasing but randomly drawn amounts of data correspond with ever increasing sampled diversity is called a subsampling curve. A rarefaction curve (Sanders 1968; Raup 1975; Tipper 1979) is a special case of the latter (see below).

The "data" can be of many kinds. Analyses of accumulation curves often contrast the number of historical years, career years, publications, or observations with the cumulative number of taxa found (e.g., Alroy 2002; Tarver et al. 2007; Benton 2008). For example, Fig. 2.2 uses the number of published occurrences as a proxy for collecting effort. Meanwhile, a subsampling curve might focus on the number of fossil collections (e.g., Alroy 1996, 1998, 2000b) or individual occurrences of taxa within collections (e.g., Miller and Foote 1996; Fig. 2.3).

The central problem in the literature is that we might be at different points on the x-axis of the curve in different geological time intervals. This problem is broadly called sampling intensity bias, as opposed to just any sampling bias. The very fact that subsampling curves do exist and do climb slowly means that whenever sampling effort varies through time, the shape of a raw diversity curve must be biased and may even include little or no biological information.

Now, being on two different points on the same curve is bad enough. Even worse, it's possible that the curves in each time interval have different slopes. If so, one interval with a steep curve (= high diversity) may be so poorly sampled that another with a shallow curve (= low diversity) may actually include more sampled taxa. Worse still, the curves might cross. Curve crossing is particularly common in accumulation curves (e.g., Figs. 2.1, 2.2). How are we to know which time interval is really more diverse if the order has switched historically as more effort has been expended?

Worst of all, we cannot make these worries go away. The reason is that we only really care about cases where the taxon counts either (1) arise from different subsampling curves or (2) fall at different points on the same curve. If the counts didn't meet that description, the temporal trend in diversity would be entirely flat and there would be nothing to talk about.

Uniform subsampling methods.—To start with, let's assume that the entire goal of subsampling is to make sure that sample size is comparable in each time interval, i.e., that uniform sampling is accurate sampling (in the sense of being statistically unbiased). This assumption is false but almost all of the literature now makes it, so bear with me.

Regardless of assumptions, all subsampling methods do the same thing when they are used to make diversity curves. You start with a set of "items" that include taxa. You set a uniform sampling "quota" to be used in all the time intervals. You grab one item, count the taxa, grab another, etc., climbing up the x-axis of Fig. 2.3 until you reach the quota (the fact that you are sampling up instead of down makes the term "subsampling" intrinsically confusing). Draws are without replacement. You do this repeatedly and take an average.

Two subsampling methods are particularly easy to understand. The first is rarefaction in the strict sense, which is most often used by ecologists when the randomly drawn "items" are specimens (Sanders 1968). The other assumes that the items are entire collections of specimens, and is either called by-list subsampling (Alroy 1996, 1998) or sample-based rarefaction (Shinozaki 1963; Chiarucci et al. 2008). So, in the first case one might always draw 1000 specimens; in the second one might always draw 10 collections.

These particular methods were not exactly used in the two papers that reignited the sampling debate in paleobiology (Alroy 1996, which briefly presented a new method, and Miller and Foote 1996). Both papers assumed that specimen counts are not available, so one has to count occurrences instead (again, a collection with ten taxa has ten occurrences and who knows how many specimens).

Miller and Foote (1996) directly rarefied the occurrence data, drawing them randomly and independently. Paleobiologists now call this method "classical" rarefaction (CR) even though the items are occurrences instead of specimens.

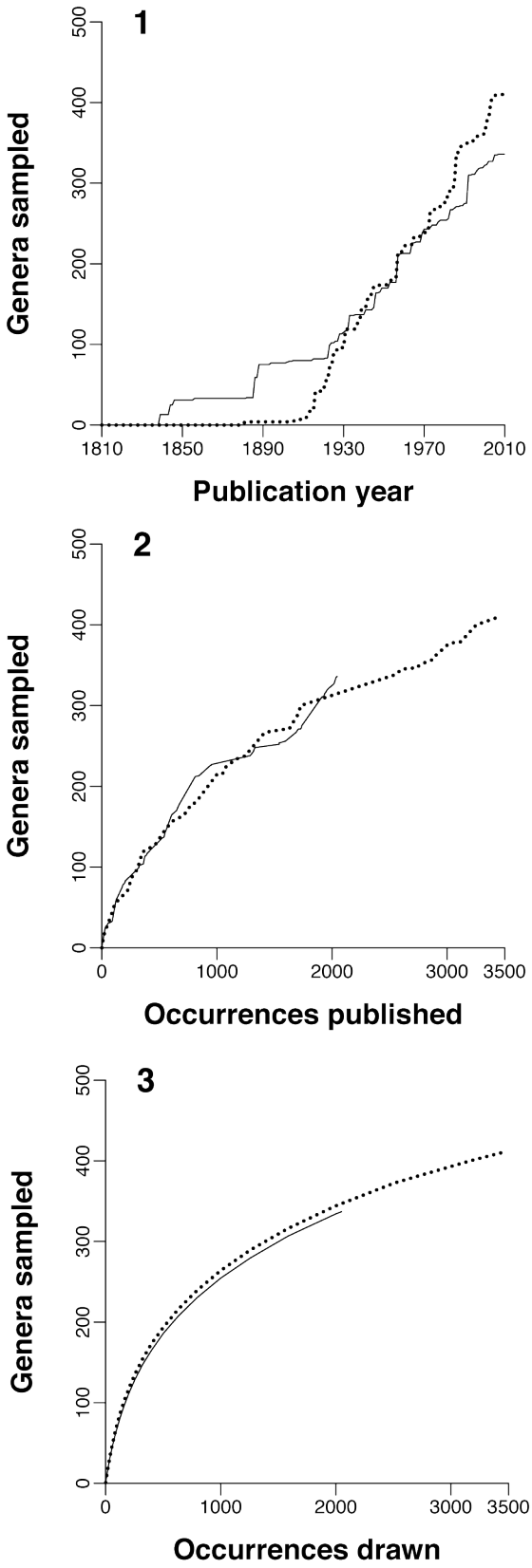

FIGURE *2–1*, *2*, *3*. Curves relating sample size to sampled genus-level diversity in bivalves. Conventional data sifting criteria (Alroy et al. 2008; Alroy 2010) are used (e.g., collections from unlithified or easily sieved sediments are excluded). Solid gray lines show Late Eocene (= Priabonian) data (up to 2050 occurrences, 435 collections, and 337 genera); dotted lines show Oligocene data (up to 3435 occurrences, 654 collections, and 410 genera). *1*, Accumulation curves relating diversity to year of publication. *2*, Accumulation curves relating diversity to the total number of collections described through historical time. *3*, By-collection subsampling curves relating diversity to the number of randomly drawn fossil collections.

Meanwhile, Alroy (1996, 1998) drew collections, so in some sense the analyses were sample-based. However, in those papers I counted occurrences to define the quota (i.e., occurrences-weighted subsampling or OW). Given the same data set and the same occurrence quota, on average each method would draw almost exactly the same number of occurrences and taxa in each interval (but not at very low sampling levels). However, OW would draw from far fewer collections because CR can draw from as many different, unrelated collections as it draws occurrences.

Suppose there are two intervals, A and B, with the following numbers of occurrences per collection: 4, 5, and 6 in A, and 1, 2, 3, and 4 in B. Because the totals are 15 and 10, the highest feasible quota is 10. With simple rarefaction, 10 occurrences out of 15 would almost always hit all three of A's collections at least once. However, any draw of two collections by OW would fill the quota immediately. So, the average number of collections drawn would be consistently lower (two vs. three). Just as importantly, two is half as much as four, the number both methods need to reach the quota in the second interval.

Now suppose we draw collections, count collections, and do not count occurrences (unweighted subsampling or UW, i.e., what ecologists call sample-based rarefaction in the strict sense). In this case the highest feasible quota is three collections. That quota gets us all of A's 15 occurrences, but on average 3/4 of 10 = 7.5 of B's occurrences.

Summary: so far we have looked at three methods, and they not only work differently but yield different expected counts. They are:

(1) Subsample occurrences individually (CR).

(2) Subsample collections but count occurrences (OW). Same expected numbers of occurrences and taxa, but fewer collections.

(3) Both subsample and count collections (UW). Different numbers of everything.

It also happens that the variance in sampled diversity is much higher if you draw whole collections. The reason is that by drawing collections one could fill the quota either with many small, boring collections or a few large, diverse collections, and both things will happen commonly.

For example, suppose there are nine collections each only including species X and a tenth including X and four other species. If entire collections are drawn and the quota is five occurrences, 10% of the time the quota will be filled immediately by drawing the one large collection, but that collection will never be drawn 50% of the time (so only one species will be sampled). By contrast, by drawing isolated occurrences with simple rarefaction one will get all five species only about 4% of the time and a single species 13% of the time. Thus, classical rarefaction yields narrow confidence intervals because it assumes an unnatural sampling process: examining some taxa from each collection and ignoring the others, instead of examining entire collections at once.

A note on terminology.—Some of the preceding may be unclear because I go back and forth from discussing "rarefaction" and "subsampling," even when explaining some of the same methods. Ecologists do not use the term "subsampling" for any of these methods: they categorize everything as either accumulation (which makes sense to me) or rarefaction (Gotelli and Colwell 2001). Furthermore, at different points I myself called OW a kind of "rarefaction" (Alroy 1996) and UW a kind of "standardized sampling" (Alroy 1999) – mea culpa again! – before deciding that all of these methods involve "subsampling" and that rarefaction was a special case (Alroy 2000b).

My reasoning was that Sanders (1968), Raup (1975), Tipper (1979), and many others all used "rarefaction" to refer to an algorithm or equation that assumed draws of items were independent and not packaged into collections. I also can't find any evidence that ecologists used the term "sample-based rarefaction" before Gotelli and Colwell (2001) popularized it, although some earlier literature called it rarefaction of some kind (e.g., Smith et al. 1985).

Putting priority aside, the term "rarefaction" is obfuscatory: the textbook definition is that it has to do with making something lighter by decreasing its density, whereas "subsampling" has a self-evident meaning and is broadly used in the statistical literature. It's also a good complement to "accumulation." Finally, I'd venture that "sample-based" doesn't really get across the idea that draws are of entire samples.

Obviously, if we want to talk to ecologists we may have no choice but to speak of "sample-based rarefaction." But if paleobiologists want to talk to anyone else, including each other, "subsampling" is a lot more clear. One way or another, the term "by-list subsampling" I tried to popularize (e.g., Alroy 2000b; Alroy et al. 2001) will have to go, because there's clear agreement now that what I called "lists" really should be called "collections" or "samples." To sum up, I think that a term such as "by-collection subsampling" is about as clear as we're going to get, and it's no less unwieldy than "sample-based rarefaction."

The collection size problem.—Let's look again at the two hypothetical time intervals. The mean number of taxa is $(4 + 5 + 6)/3 = 5$ in the first, and $(1 + 2 + 3 + 4)/4 = 2.5$ in the second. There are two obvious explanations for the drop:

(1) True alpha (local scale) diversity fell. Here, alpha is the absolute number of taxa in the entire species pool of a given collection.

(2) The average number of specimens in each collection fell.

Without actual specimen counts we cannot answer the question definitively. Putting CR and its dodgy assumptions aside, here is what we should do:

(1) If alpha changes but collection size doesn't, one should draw collections (UW).

(2) If alpha is constant but collection size changes, one could weight collections by occurrences (OW) and hope that the true, unknown specimen count is a always simple multiple of the occurrence count. For example, one could hope that a set of collections totalling 200 occurrences would include twice as many specimens as a set totalling 100.

Now, it only really makes sense to assume that collection size varies. That's why we rarefy individual collections when we have abundance data! Furthermore, the huge variation of collection-level diversity in real paleontological data sets suggests that collection size does vary. Intuitively, then, if the goal is uniform sampling then some method related to OW would seem preferable.

The problem is that OW's assumption of a linear occurrences-to-specimens ratio is wrong. I already have mentioned there is an asymptotic relationship between the number of specimens drawn and the number of taxa seen (for examples, see Alroy 2000b and Bush et al. 2004). Therefore, if you reverse the axes you will see an upwardly curved relationship between the number of occurrences drawn and the number of specimens represented by these specimens.

So, how can we guess the number of specimens from the number of occurrences more reasonably? For mammals, I argued from a few abundance data sets that the rarefaction curve for the typical collection mostly followed a simple power law, or log-log linear function, with a slope that might change through time (Alroy 2000b). A power of 2.0 seemed to fit well in some key cases (as in the raw data of Bush et al. 2004): the number of specimens roughly equalled the square of the number of occurrences. Therefore, I suggested that while drawing collections, the sum of these squared values should be counted and used to set the quota (Alroy 2000b).

Like OW, the occurrences-squared or O2W method assumes that alpha diversity does not change. More occurrences means more specimens, period. Here are the specimen count estimates for the two intervals: $16 + 25 + 36 = 77$ and $1 + 4 + 9 + 16 = 30$. Suffice it to say that a 77:30 O2 ratio is not the same as a 3:4 collection ratio or 15:10 occurrence ratio. Use this method (with one exponent or another) and you will draw a different amount of data in each interval than you would have drawn using UW or OW.

The alpha and beta problems.—As mentioned, Bambach (1977), Bush and Bambach (2004), Kowalewski et al. (2006), Alroy et al. (2008), and others all argued that, in fact, marine invertebrate alpha diversity has changed a lot through time, and specifically was much higher in the Cenozoic than previously. Whether this pattern is mostly or only partially explained by better preservation in the Cenozoic remains an open question (e.g., Hendy 2009; Sessa et al. 2009), but take it on face value for now.

Perhaps, then, the rarefaction curves for individual collections vary through time (just the same as saying that alpha diversity does). If so, the O2W rule (specimens = occurrences squared) does not always hold. Instead, you should vary the slope of the assumed power law: maybe it is 2.0 sometimes, but drops as low as (say) 1.4 in others. I pointed out this possibility (Alroy 2000b) and tried to deal with it using a very indirect method I won't detail and didn't use again. Instead, I used O2W in two of the eight analyses presented in Alroy et al. (2001). Later (Alroy et al. 2008), I went back to the problem and proposed a simple way to calibrate the slope of the occurrence weights in each time interval by looking at rarefaction curves for the individual collections that do have abundance data.

Another problem, which was noted by Bush et al. (2004), is that methods involving occurrence weighting can fill quotas with a few big collections and a lot of little ones. As a result, if there are some really big collections O2W in particular might create strong geographical clustering of the data, and therefore might underestimate global diversity. I went to a lot of trouble to handle this problem with a simple method called inverse weighting (Alroy et al. 2008). Because this method is discussed elsewhere and because I don't use it now, I won't go into detail about it.

Uniform sampling vs. fair sampling.—So far so good. We know the problem and its solution: sampled diversity varies because sample size varies, so we must make sample size uniform. That means drawing a uniform number of this, that, or the other thing and hoping that doing so indirectly yields exactly so-and-so many specimens in each time interval.

Now, hold on. Let's remember that what we really want is fair sampling, which is not by definition uniform sampling, and "fair" should mean "accurate." The definition of an accurate curve is not very complicated: it should have exactly the same shape as the historical curve of everything that ever existed. That is, the standardized counts should be scaled down from the changing size of the entire species pool by some fixed ratio.

However, drawing a fixed, uniform number of things will never give you fair sampling in that sense.

Here's a simple example: suppose you always draw one specimen. Period. Your "standardized" curve will be completely flat (one specimen = one taxon) and therefore completely inaccurate.

Here's another example. Suppose that you have two intervals A and B. In A, there really truly is only one species in the world. In B, there are *billions* of species, all rare, so you have to look at *millions* of specimens to get even one species twice. Now if you draw X times more specimens, you will get X times more species in B but keep getting only the same one in A. So, standardization will tell you nothing about true relative diversity.

We shouldn't be surprised by the failure of uniform sampling because what it does simply isn't intuitive: ignore rarity and commonness and only count items and kinds of items. Intuitively, when there are just a few kinds of things we should not have to sample too hard; when there are tons of them and we keep finding more as we collect, we should just keep collecting. Everyone does this when they're not forcing themselves to collect systematically. But analytical paleobiologists have ignored this intuition, and that's why many field paleoecologists find sampling standardization so unsatisfying.

Shareholder quorum sampling.—Here is my solution (Alroy 2010).

Another thought experiment: suppose one species pool A has 10 equally frequent species (i.e., each one has a frequency $f = 0.1$). If you draw two specimens, on average you get 1.9 species in total because the second draw will give you a different one 90% of the time.

Now suppose we create a second pool B by adding 10 more species that are equally good competitors and therefore equally frequent. We should also expect real data to change in an evenhanded way because there is no obvious, unavoidable reason that new species should be relatively common or rare. Now $f = 0.05$ for all 20 species, and now two drawn specimens will on average yield 1.95 species. This number is wrong. Instead, an accurate method would have drawn exactly twice as many species (3.8).

Suppose, however, that we treat each species as a "shareholder" whose "share" is its frequency. As we randomly draw specimens, we ignore the specimen count and focus on the "share" that is represented so far. Each species' full share is considered to be represented the very first time it is sampled. You stop when a certain fraction of shares (not specimens, and not species) is represented by at least one species at the "shareholder's meeting." When that happens you have a "quorum."

Suppose the desired quorum (sum of represented frequencies) is 0.2. Clearly, with 10 species and $f = 0.1$ you will always get two species when you reach that quorum. With 20 and $f = 0.05$, you will always get four. Four is twice as many, and that is exactly what you want because B is truly twice as diverse as A. And no matter what the quorum, you will still have perfect accuracy because the 1:2 ratio of real diversity in A and B will match the 1:2 ratio of sampled diversity. The same thing can be seen in more complicated examples (Fig. 3).

Problem solved.

Good's u *and the evenness problem*.—Not exactly. We still have to deal with one important technical issue and a pair of headaches. The issue, and in fact the only reason we need to have a quorum subsampling method in the first place, is that we can't know the true underlying frequencies until we get every last species in the pool. If, say, frequencies are roughly equal and the pool has 100 species, then on average $f \sim 0.01$. However, if our entire fossil data set has only 10 species, we will think that $f \sim 0.1$ for each of them. If we set the desired quorum q to 0.2, we will get two species and stop when we really needed to draw 20.

Before continuing, I need to introduce the term "coverage" (Good 1953; Colwell and Coddington 1994), which is not to be confused with the "coverage probability" of a confidence interval in classic statistics. It means the proportion of the entire frequency distribution represented by the species you have found so far. The quorum is a certain amount of coverage: if $q = 0.40$, you stop sampling when you have "covered" 40% of the distribution.

We can now rephrase the problem that motivates the method: our entire actual sample does not cover the entire actual distribution. It might cover 90%, 95%, or even 99.9%, but not everything. Therefore, a species that has a frequency of 0.1 in our real data most likely "covers" less than this amount of the entire true distribution.

Very fortunately, the problem of overestimating frequencies in a finite data set was solved a lifetime ago by Good (1953), who came up with a jiffy formula

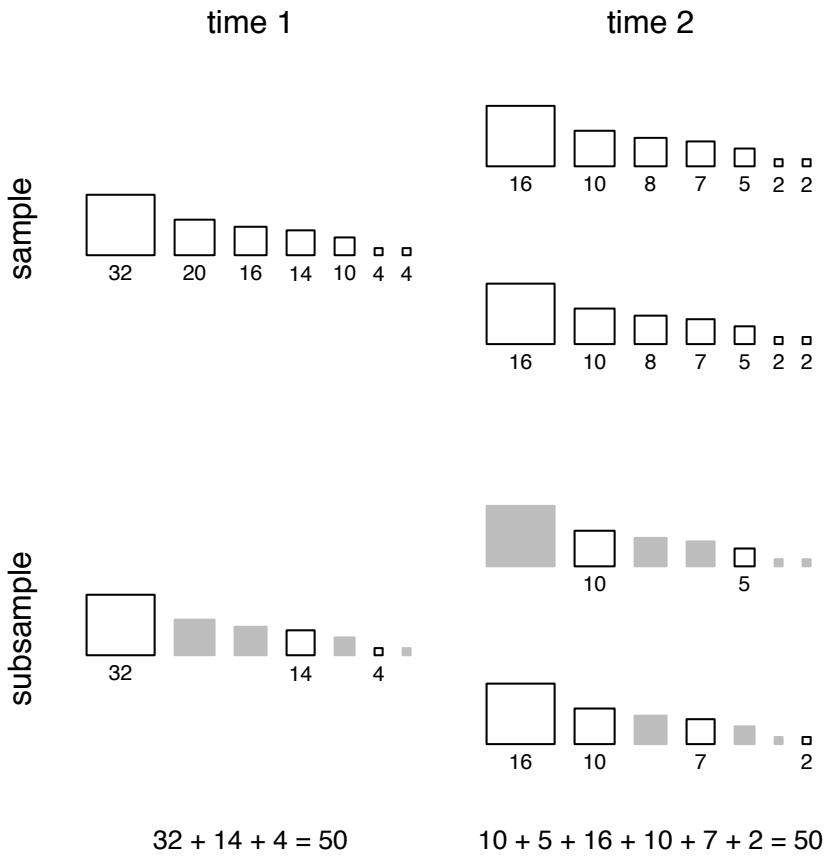

FIGURE *3*. Effect of changing species pool size on shareholder quorum subsampling. Squares = taxa; numbers = percentage frequencies. All taxa in the left column (representing one time interval) are present in the right column (representing the next), but are matched with other taxa having the same distribution of frequencies (so relative frequencies are halved). Top row shows frequencies in the raw data set (the pool is assumed to be fully covered to simplify the arithmetic, so frequencies add up to 100%). Bottom row shows taxa randomly drawn in a single subsampling trial with a quorum of 50% coverage.

for coverage:

$$u = 1 - n_1/O \quad (4)$$

where O = the number of observations (say, specimens or occurrences) and n_1 = the number of singletons (species represented by exactly one observation). Suppose you have 100 specimens and 10 singletons. Good would say that your coverage is 1 - 10/100 = 0.9. Therefore, if there are 20 specimens of species X, its observed share is 0.2, but its true share is most likely 0.2 x 0.9 = 0.18. In our situation, we would simply substitute occurrence counts for specimen counts to obtain this figure.

Again, problem solved. Every time we analyze a time interval, we compute u and divide it into q (Alroy 2010). Suppose interval A has really bad coverage and its u = 0.4, and B has pretty good coverage and its u = 0.8. The highest we can make q is 0.4. The reason is that if q = 0.4, in interval A q/u = 0.4/0.4 = 1 = sample everything, and you can't drawn more than everything. If we do settle on q = 0.4, then in interval A we will have an effective quorum level of q/u = 0.4/0.8 = 0.5. In other words, we will stop drawing data when we see species representing 50% of the raw frequency distribution (which could happen almost immediately if some species are very common, or only after exhaustive sampling if there are many species and the abundance

distribution is very even).

The evenness problem.—As you might have guessed, the key assumption of shareholder quorum sampling (SQS) is that species are added to the species pool or subtracted from it in a way that is random with respect to their frequencies. If so, simple arithmetic guarantees that the method will work every time.

Obviously, however, this assumption is sometimes wrong. If, for example, super-abundant species join or leave the abundance distribution, the distribution will respectively become less or more even. The consequences could be serious. Suppose you add one big "weed" species with an occurrence frequency of 0.5. That will blow out your quorum of 0.4 right away, even if the pool size hasn't changed much.

To make a long story short, based on unpublished simulation analyses I think we can mitigate this headache very simply: just ignore the one most common taxon. The frequency of the most common thing actually has a fancy name in ecology, namely, the Berger-Parker index *d* (Berger and Parker 1970; May 1975). So, my approach (Alroy 2010) is to discount *d* when you compute coverage: if o_1 is the dominant taxon's observation count, then:

$$u = 1 - n_1/(O - o_1) \quad (5)$$

And when the dominant taxon is encountered during subsampling, you don't count its frequency towards the quorum but you do add one to the diversity estimate.

The "new species of X from Y" problem.—The second headache has to do with the very nature of scientific literature: the whole point of publishing is to describe new things. Therefore, literature-based occurrence data sets are not a random sample of the fossil record, but a sample combining hopefully random occurrences of old things with definitely nonrandom occurrences of new things. Many of these taxa are technically not singletons because multiple collections yielding them are listed in the original description (and these collections typically are all from the same stratigraphic horizon and the same small geographic area).

The practical problem is that new-taxon occurrences cause diversity estimates to rise and rise as more literature comes into a data set. This bias affects all subsampling methods, not just SQS. Its presence can be demonstrated by analyzing any large data set and then reanalyzing it after restricting the data to collections published in a random, uniform draw of references (Alroy et al. 2008). The analysis using less references will yield a lower diversity curve every time, even if the sampling level is the same.

Statistically speaking, then, only the old, boring occurrences are reliable. It would be tempting to simply throw out the other ones, but that would reduce the number of singletons to zero and thereby make coverage appear to be 100%. Instead, the solution (Alroy 2010) is to reformulate Good's equation by counting occurrences of single-publication taxa (p_1) instead of single-collection taxa (n_1):

$$u = 1 - p_1/O \quad (6)$$

Including the other tweak that involves dominant taxa, the full equation is:

$$u = 1 - p_1/(O - o_1) \quad (7)$$

Or:

$$u = (O - o_1 - p_1)/(O - o_1) \quad (8)$$

In a small data set, the vagaries of the literature can create a seemingly related but actually opposed problem: a large fraction of the occurrences may come from a single large collection (or a very broadly defined one). Such a collection would inflate p_1 and thereby push down *u*, but the existence of a very large collection should indicate good coverage (or at least not indicate bad coverage). Alroy (2010) addressed this problem by restricting the single-publication taxon count (p_1) to taxa not found in the largest collection.

The taxonomic coverage problem.—It gets worse. There is yet another problem with literature data: systematicaly uneven coverage, especially of different major taxa that may have different levels of diversity and ecological abundance and preservability. For example, Paleozoic brachiopods were fairly diverse and very abundant and preservable; Cenozoic gastropods were very diverse and abundant and preservable. If nothing is done about this, a Phanerozoic "marine" curve will more or less become a curve of brachiopods, gastropods, etc., and will be too high at the end (i.e.,

during the Cenozoic, a.k.a. the Age of Gastropods and Bivalves).

Alroy (2010) solved this problem simply by computing separate diversity curves for each major group and then summing them to get a master curve. Likewise, Alroy (in press) tried to remove problems with uneven geographic coverage by computing and summing separate diversity curves for northern, southern, and low-latitude regions. That kind of maneuver can only be a good thing when there are big, systematic differences between groups (or geographic regions, or environments, or what have you).

The problem is that many data sets are simply too small to be split in any way. Fortunately, a different approach also more or less solves the problem: draw data from as many different references as possible. Alroy (in press) did so with a "throwback" algorithm. Every time a collection was drawn during subsampling, it was thrown back into the sampling pool with a probability inversely proportional to the number of collections yielded by the same publication. So, for example, if you drew a collection that came from a reference that included 10 collections, that collection would be thrown back (temporarily) 1 - 1/10 = 90% of the time.

Both strategies yielded pretty much the same Phanerozoic diversity curve. If that hadn't happened, something would have been seriously wrong. Also, if the quorum level had been set close to the total available coverage in many time intervals, none of this monkeying around would have made any difference because most available collections would have had to have been drawn anyway. That's one reason it's definitely a good idea to look at results based on multiple quorum levels.

So what?—That was an awfully long explanation of a very strange method. But isn't it just another tweak? The answer is no, because shareholder quorum subsampling is not just another item quota subsampling algorithm. There is a big conceptual leap between thinking that fair sampling is uniform sampling (i.e., it should yield a fixed amount of data) and thinking that fair sampling is accurate sampling (i.e., it should yield a fixed proportion of the species pool). Some will disagree, but I do think this difference qualifies as a paradigm shift.

It would still be easy to think that SQS somehow uses a quota, so let me make things really clear: SQS does not draw a fixed number or fraction of data items or taxa. Nothing is fixed. The number of things and number of different things can vary all over the place.

Empirically, though, the difference between item quota subsampling and quorum subsampling is not always very great. The new shareholder quorum curve for the Phanerozoic marine data set (Alroy 2010) reproduces a lot of important features that are seen with quotas (Alroy et al. 2008). All of the key transitions appear in both curves and many fine-scale features are the same. Most importantly, there is still no massive exponential post-Paleozoic radiation. In fact, there is still no large net change through the Cenozoic.

The new curve does show a jump across the K-T boundary that is tied to a big radiation of gastropods. It also has a higher amplitude in other places: the mid-Devonian, Permo-Triassic, and Triassic-Jurassic crashes are all much steeper and the mid-Jurassic radiation is more dramatic. My bet is that most of these differences are real. But one way or another, we should be way past the point of needing to argue about whether standardization in general is a good thing. The only question is what method to use. It is also now clear that a compounded series of sampling and counting biases are responsible for the big exponential radiations drawn out of Sepkoski's data by certain protocols (e.g., Benton 1995; Bambach 1999), if not by others (e.g., Sepkoski 1997).

ORIGINATION AND EXTINCTION RATES

Extinction events have been a major focus of the literature ever since Cuvier convinced his colleagues that extinction does happen with his description of the American mastodon in 1796, and later with his demonstration in 1811 that entirely distinct faunas had occupied the Paris basin through the Tertiary (which of course suggested mass extinctions and replacements by the hand of God: Rudwick 1998).

The first substantial use of extinction rate data was by Lyell (1830), who defined the Eocene, Miocene, and Pliocene epochs on the basis of percentages of extinct molluscan genera in different assemblages. For example, 176/1021 species in the (Miocene) formations of several countries were extant but only 38/1122 species in the (Eocene) Paris basin were, so the Eocene rocks had to be much older. *Lyellian curves* (e.g., Raup and Stanley 1971; Stanley 1973) are plots showing how

such percentages decrease through time.

Lyell argued that the divisions did not have "too much importance" because their differences are heightened by gaps in the fossil record, and that more research would close the gaps and lead to further subdivisions. So, Cuvier and Lyell differed completely on whether (1) the faunal overturns were rapid or (2) the gaps were large. These alternatives are still debated (Foote 2005; Alroy 2008).

Darwin (1859) had little to say about extinction in the fossil record other than to echo Lyell's views, and most work on extinction was descriptive and narrowly focused for the next century. One of the key exceptions included Phillips (1860), who presented a hand-drawn Phanerozoic diversity curve that recognized the Permo-Triassic and K-T mass extinctions (which he used to set off what he called the Palaeozoic, Mesozoic, and Caenozoic).

Even Simpson (1944) was much more interested in character evolution, speciation, and the origins of major groups than extinction, and did not view it as a major factor needed to explain biological history. However, he did examine taxonomic duration data to show that mammals evolve (or more accurately go extinct) faster than bivalves. He also (Simpson 1952) published an important paper on origination rates (not extinction rates) that argued against physical environmental factors as a driver of "explosive evolution" (adaptive radiations).

Exponential decay models and survivorship analysis.—Quantitative analyses of turnover rates really began with Van Valen (1973), who argued that extinction could be modelled as an exponential decay process ("Van Valen's law"). Taking off from Simpson's 1944 analysis and a later paper by Kurtén (1960), he looked at cumulative durations of taxa plotted on a log-linear scale (i.e., a *survivorship analysis*).

The x-axis in such a plot is time since each taxon's origination, whenever that was, and the y-axis is the log of the number of taxa still alive after that amount of time. If there is a straight-line falloff in such a distribution then there is a continuous process of extinction that is non-selective (random) with respect to taxon age. The slope of the plot is equal to the decay rate, i.e., the per-capita instantaneous extinction rate.

There are two similar ways to analyze survivorship data. As I mentioned, each point can show the number of things still alive after some amount of time (e.g., people still alive at age 44). Alternatively, the points could show the number of things that disappear during uniformly spaced time intervals after originations (e.g., people dying during the years they were 44, 45, etc.). Both counts should decay exponentially, and with the same rate. The difference is that death counts are small, so they are noisy (bad) but statistically independent of each other (good). If you want to get the extinction rate with a linear regression, you should therefore use the still-alive counts if you have little data and the death counts if you have a lot of data.

Lyellian curve analysis.—Apart from adding up durations, there are two other general strategies for getting rates that are both useful and both underused in the paleobiological literature.

The first is to look at trends in Lyellian curves, which is basically a special kind of survivorship analysis because in both cases you look at the proportion of taxa surviving some amount of time. However, here the x-axis is "time before the Recent" instead of "time after origination." Also, each data point represents all taxa found in a *single* time interval instead of all taxa that ever reached a given age. Likewise, each taxon is implicitly examined to see if it continued up to a *single* point in time (now) instead of being examined to see if it went up to different points.

Despite all of that, the logic is basically the same: if turnover is random and continuous, the curve's shape will be log-linear. Therefore, if you take proportions, log them, and fit a line, the slope will again define the extinction rate. The only really important procedural difference is that raw counts can't be used because diversity changes through time. Therefore, you must convert the values to proportions (or percentages) before you log them.

There are two big advantages to Lyellian analysis. First, sampling of extant taxa with good fossil records is effectively 100% in the Recent (that explains why very few taxa are ever found alive after first being described from fossils). Therefore, you can really trust the percentages and really ignore how sampling might work in the fossil record.

Second, although the data points can represent all taxa ranging through an interval, they really should represent all taxa found as fossils in an interval – or even in a specific fossil collection (which is the way

Lyell himself did it). Because each fossil collection is an independently drawn batch of information, the resulting data points are statistically independent (you can't predict anything about one percentage even if you know all the others). Thus, at least in some cases Lyellian plots can combine the advantages of statistical independence (as with death counts) and large sample sizes (as with cumulative durations). The only problem is that the method only works for large extant groups.

Cohort analysis.—Second, in a *cohort analysis* you can extract the extinction rate from a cumulative duration plot for the taxa that originated in a single interval (Raup 1978). By contrast, each taxon in a Van Valen plot starts when it starts: the cohorts are all lumped together, and you only look at the durations. So, here you focus on one group of taxa starting at one point (say, the Pliocene cohort), as in a Lyellian analysis, but you do look at the taxa again and again as you go forward in time, as in a standard survivorship analysis.

Cohort analysis is useful for several things other than just defining an average rate. First, you can spot mass extinctions very easily in cohort data, which is almost impossible in a Van Valen plot. Of course, you can do the same thing with any time series of extinction rates. Second, cohorts can have different average extinction rates and different responses to mass extinctions, which is likely to be biologically interesting (e.g., Foote 1988). Finally, Raup (1978) figured out a way to estimate species-level extinction and speciation rates from genus-level cohort data, and believable species-level rates would be wonderful to have. However, Raup's speciation rate estimates are always lower than extinction rates (not higher!) because speciations leading to new genera are not considered (Gilinsky and Good 1991). His method is also much more informative about net rates of diversification than about particular speciation and extinction rates (Foote 1988). As a result, few paleontologists now use Raup's particular method or any other kind of survivorship analysis.

The Marine Biological Laboratory (MBL) simulation.—Almost at the same time as Van Valen's survivorship analysis paper, a group of paleontologists joined by ecologist Dan Simberloff met at Woods Hole and brainstormed a computer simulation model. The resulting paper (Raup et al. 1973) got everyone thinking about evolution as a mathematical process that was governed by general underlying laws but had random individual outcomes. Although not listed as an author, Gould's graduate student Jack Sepkoski wrote the actual FORTRAN code they used (Sepkoski 2005: he was finally credited in Gould et al. [1977], one of several followups).

Raup himself later published most of the key work on birth-death = speciation-extinction processes as they pertain to the fossil record (e.g., Raup 1978, 1985). His equations can be used to infer all sorts of interesting things, such as the probability that a clade will survive a certain amount of time. However, we will not go into further details about such models here. Our interest is more empirical: extracting rates from real data.

Clade-specific rates.—There are two categories of rates one might want to get: average rates for individual taxonomic groups and rates for individual time intervals.

The Woods Hole simulation assumed that the net diversification rate for a group would start out very high and then fall to zero as the group's diversity rose to a limit (i.e., an equilibrium level: Sepkoski 1978). Group-specific rate models of this kind became an immediate focus. Extinction rates are relatively easy to get (see above), but speciation rates are much harder to infer without a phylogeny (and still very hard to infer even with a complete phylogeny of living species, for what that's worth: Quental and Marshall 2009).

Suppose we call extinction and speciation λ and μ (following Raup 1985). Stanley (1975) tried to solve the problem of estimating λ by assuming that (1) you have already gotten the extinction rate μ from something like a Van Valen plot, (2) the group is extant, (3) the count of living species N_R is accurate, (4) you know the group's origin time (time 0, so right now is time t), and (5) the net diversification rate $r = \lambda - \mu$ is constant, meaning that growth is purely exponential (so current diversity is equal to $N_0 e^{rt}$). If so, then because:

$$N_R = N_0 e^{rt} = N_0 e^{(\lambda-\mu)t} \quad (9)$$

and because assumptions 1 through 5 together mean that we know N_R, N_0 (= 1 by definition), r, t, and μ, we can now get λ:

$$\lambda = \ln(N_R/N_0)/t + \mu \quad (10)$$

Obviously, there are big problems with the assumptions!

Traditional per-interval turnover rates.—Surprisingly, analytical paleobiologists made no connection between birth-death models and turnover rate time series throughout the 70s, 80s, and 90s. Instead, they typically used one of three measures: raw counts of events E; percentages, i.e., E/N where N is the standing diversity level (usually based on the total range-through count, N_r); and "rates" that were simply the raw counts E divided by time interval durations Myr (which are not at all the same as instantaneous rates).

For example, Simpson (1952) used events-per-Myr origination rates. Webb (1969) and Stanley (1973) used the same rates in early studies of mammal diversity dynamics and Phanerozoic bivalve origination patterns. So did Raup and Sepkoski (1982) in their famous "decline of extinction rates" paper.

However, Raup and Sepkoski's equally famous "periodicity of mass extinctions" paper (1984) instead used extinction percentages. They discussed the "rates" they had used before, and concluded that normalizing by interval duration was a bad idea because raw extinction counts are not correlated with interval durations (virtually the same argument is made by Alroy [2008]).

Sepkoski (1978, 1979, 1984) was very clear on the idea of instantaneous rates a la Stanley (1975) in his famous "kinetic model" papers. All of his models were developed in this framework. Nonetheless, in all three papers he used events-per-Myr rates whenever he showed real data. As late as 1991 Raup himself was publishing per-Myr percentage rates equal to E/(N Myr) (the fourth and last of the obvious combinations of event counts, standing diversity levels, and time interval durations).

One important exception to the use of simple ratios is Gilinsky and Good (1991). They introduced a method for estimating turnover rates that was based explicitly on a branching model and got its parameters with a maximum likelihood algorithm. However, their method assumed an underlying process that worked in discrete time steps, which doesn't make much biological sense, and it produced rate estimates of zero for multiple time intervals, which I believe makes even less sense.

Foote's instantaneous per-time interval turnover rates.—Not only did most time series studies ignore the instantaneous rate literature, but Raup and Sepkoski (1984) turned out to be prophetic: not just some but all of the per-Myr equations are likely to suffer from a negative correlation between rates and interval durations (Foote 1994). At the time, the best option open to Foote (1994) was to work with simple percentages (= E/N), which creates other problems.

Foote (1999) was the first to make real progress in this area by finally seeing the connection between branching models and empirical turnover formulas. In an appendix to a monograph he presented the first (paleontological) equations for computing time-interval-by-time-interval instantaneous rates, as opposed to average rates for whole taxonomic groups (similar things had been done by population ecologists). He later wrote a full paper on the topic (Foote 2000).

Foote's innovation was to treat each rate as a mini-cohort analysis. The "cohort" was the number of taxa present at the base of the bin, i.e., the boundary crosser counts. In the notation given earlier, BC is equal to $N_{r+} + N_{r-} + N_b$ (eqn. 2). In Foote's notation, $N_{r+} + N_{r-} = N_{bt}$ (number crossing both the bottom and top). His N_b and N_t (as used in 1999) are the same as what we defined earlier (Fig. 1). So, $BC = N_{bt} + N_b$.

Foote (1999) then looked at how much of this BC cohort survives to the end of the bin, which of course is N_{bt}. Putting aside variation in interval lengths, the survival rate is just the log ratio $\ln(N_{bt}/[N_{bt} + N_b])$. Because this number is an exponential decay coefficient, we can think of it as the extinction rate: the lower it is the less the chance of survival and the greater the chance of extinction. And because this number is negative, we can invert it to get a positive number that is more intuitive:

$$\mu = \ln([N_{bt} + N_b]/N_{bt}). \quad (11)$$

A similar expression for λ is easy to figure out. And although this is generally a bad idea, interval lengths are easy to put in (just divide μ or λ by interval duration: Foote 1999). The Foote equations were later rederived and explained in somewhat different terms (Alroy 2000b).

Foote's "optimized" maximum likelihood turnover rates.—Foote (2001, 2005) next built an entirely new method of estimating origination, extinction, and preservation rates around cohort analysis (if not Raup's

particular genus vs. species method). This approach involves fitting a maximum likelihood model to survivorship data for all the possible cohorts in a time series. All three kinds of parameters are assumed to have different values in different time intervals.

Surprisingly, with Sepkoski's genus-level data Foote's method actually increases the variation in rates and produces many "most likely" rates of zero (Foote 2005). The standard "Foote rates" discussed above (which he calls "face-value" rates) do not have this property, and neither do the new rate equations discussed below.

Foote (2007) argued that zero rates should be taken seriously, in part because the Signor-Lipps bias effectively averages adjacent rates together (and therefore smooths them out). However, I'm not sure I agree. I suspect that major biases in real data will tend to increase variance, not decrease it – especially if only a few truly large biases are present and they overwhelm the Signor-Lipps effect. I also suspect that the reason for all the zero values is allowing each of the rates to vary across all the intervals. Simpler models might show something different.

However, what I might suspect really doesn't matter, because there is little reason to use traditional range-based data now that occurrence-based data are so easily available. These data make it possible to use the new methods discussed below, which entirely get rid of problems like Signor-Lipps (Alroy 2008, 2009).

Two- and three-timer methods: basic idea.—As mentioned, the problem with the counts used by Foote (1999, 2000) is that they are based on the usual range data that have all the biases (such as edge effects) we now know are serious (Foote 2000; Alroy et al. 2008). For example, the extinction rates are likely to be too low at the end of a time series because of the Pull of the Recent (see below).

In diversity studies, the solution is to just look at what is actually sampled (sampled-in-bin diversity). By analogy, the solution for turnover rates is to only look at three counts (Fig. 4) described by Alroy (2008):

(1) Two timers, i.e., taxa sampled immediately before and within a bin, which is the same as saying before and after a bin base (N_{2t}). N_{2t} is a subset of the boundary crossers, and it was introduced earlier in the discussion of counting methods.

(2) Three timers, i.e., two timers also found in the following, third bin (N_{3t}). N_{3t} is a subset of the range-throughs.

(3) Part timers, i.e., taxa sampled immediately before and after a bin but not within it (N_{pt}). Also a subset of range-throughs.

These three fundamental counts are important because they will give us the same three parameters Foote (2001, 2005, 2007) wants to get: the sampling probability, instantaneous death rate, and instantaneous birth rate.

Because these counts ignore everything that happens outside of the focal time interval's immediate vicinity, they do not suffer from the edge, Signor-Lipps, and Pull of the Recent effects, and they suffer from rate effects only when turnover is continuous and rates are very high. They also avoid the Lagerstätte effect (one-interval sampling universe effect), like Foote's methods but unlike conventional, range-based counts and ratios. The reason is that a Lagerstätte taxon cannot be sampled twice or three times (by definition). Because of these advantages, there are good reasons to believe that the rates and sampling probabilities we are about to examine are about as unbiased as you are going to get with fossil data.

Sampling probabilities.—Before we get to the death and birth rates, let's define the part-timer sampling probability (Alroy 2008). It is the conditional probabillity of being sampled given that you are definitely present throughout the bin:

$$p_s = N_{3t}/(N_{3t} + N_{pt}) \quad (12)$$

An alternative to p_s, used by Foote (2000), is also a conditional probability involving taxa that fully range through, but it considers all full range-throughs instead of only N_{3t} and N_{pt} (which together represent all taxa sampled both immediately before and after the bin). It is called the gap percentage (Paul 1982). The ratio is:

$$N_{r+}/(N_{r+} + N_{r-}) \quad (13)$$

Again, the practical difference is that the chance of being a range-through depends on sampling probabilities in all the previous and succeeding bins. You don't want to use a number that depends on everything going on everywhere in the data set: you only care about the local sampling probability. So, this older

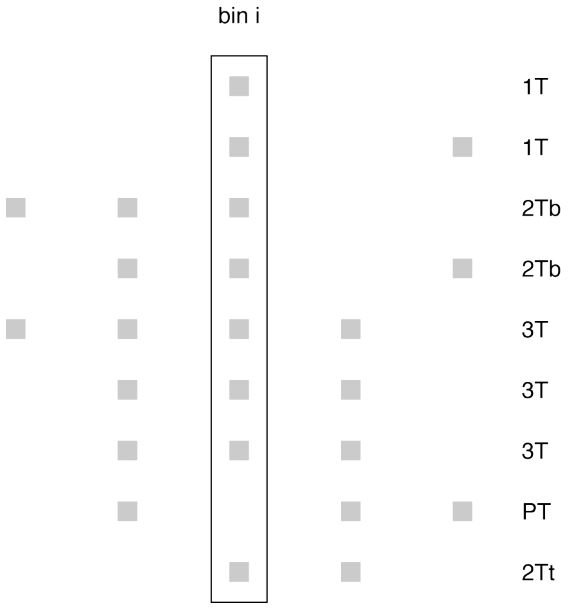

FIGURE *4*.—A taxon sampling pattern that could generate the age range data in Fig. 1. Each row shows presence-absence data for one taxon across five time intervals (bins). Black rectangle represents the third bin (*i*). 1T = one timer, 2Tb = two timer with occurrences spanning the base of i, 2Tt = two timer with occurrences spanning the top of *i*, 3T = three timer, PT = part timer.

equation is subject to all the counting method biases, unlike the new one.

However, I should point out one minor technical problem with eqn. 12 that I have already hinted at: it assumes that there is no turnover within intervals. Now, because taxa appearing or disappearing within an interval are not present all the way through it, they have reduced sampling probabilities. But these taxa are completely excluded from eqn. 12 (like eqn. 13) because it only considers taxa definitely present throughout a given interval. That means the estimate could be too high with respect to some taxa at some times.

I'm not worried for two reasons. First, there is decent evidence that most turnover actually happens at the boundaries between geological time intervals (Foote 2005, 2007; Alroy 2008, 2010). Second, if turnover rates are close to the values paleobiologists are used to seeing (say, 20 or 30% per time interval), then most taxa that ever existed in an interval either (1) were present all the way through it, or (2) had ranges spanning most of it. Nonetheless, this problem should be resolved at some point with improved equations (perhaps involving recursive computations or maximum likelihood calculations).

Three-timer turnover rates.—Let's restate Foote's basic face-value equation (1999, 2000) in a more general way. Again using the notation (and concepts) of Raup (1985), the "instantaneous" death rate is the decay constant μ (mu) of an exponential equation, where N_i is a count (of any kind) at the base of a bin called *i*; N_{i+1} is the count of the same taxa (the cohort) that make it to the top; and *t* is the elapsed time in millions of years:

$$N_{i+1} = N_i e^{\mu t} \quad (14)$$

Assume for simplicity that *t* is 1 and that μ should be a positive number (which requires switching N_i and N_{i+1}). Therefore,

$$\mu = \ln(N_i/N_{i+1}) \quad (15)$$

All we have done is restate a general version of the Foote extinction rate formula to show that it works for any kind of boundary counting method (because N_i could be anything). However, because N_{2t} and N_{3t} are pretty much unbiased, it's best to use them. N_{2t} is just the count of the cohort at the bin base, and N_{3t} is the count of the same cohort at the top.

There is a problem, however. N_{3t} is not simply the survivors of N_{2t}, but the survivors that have been sampled once again in the third bin. Just because you were sampled twice does not mean you will be sampled three times. We therefore need to multiply N_{2t} by p_s to make N_{2t} and N_{3t} comparable. That gives us:

$$\mu = \ln(p_s N_{2t,i}/N_{3t,i}) \quad (16)$$

or:

$$\mu = \ln(N_{2t,i}/N_{3t,i}) + \ln(p_s) \quad (17)$$

It's easier to understand the second version because you can think of its right-hand term as a simple correction factor.

Likewise, the birth rate λ (lambda) is what you get by thinking in reverse, with a cohort $N_{2t,i+1}$ at the top of the bin "dying" as you go backwards to the start:

$$\lambda = \ln(N_{2t,i+1}/N_{3t,i}) + \ln(p_s) \quad (18)$$

We will call these rates "three-timer" rates because they quantify the chance of becoming one.

Three-timer rates: the fine print.—It's important to note that the p_s term used in these equations can be either global (computed from total counts of all three- and part-timers found anywhere in the data set) or specific to the relevant time interval (i+1 in eqn. 17 and i-1 in eqn. 18). Generally, it makes sense to use local counts whenever they are large and global counts otherwise – even in the same time series, as I did when analyzing individual marine animal groups (Alroy 2010). Be careful with your decision.

Meanwhile, let's return to the rate bias mentioned above (which again only applies if turnover is continuous). The problem here is downwards bias in N_{3t} caused by failure to resample actually surviving two-timers in a third bin. The $\ln(p_s)$ term is supposed to fix that, but it won't work if turnover is continuous because some three timers will go extinct within this third bin and therefore won't have as many chances to be sampled.

If the sampling process is Poisson, you can show easily that the degree of bias is almost independent of the rate. For example, if the average number of sampling "hits" per taxon per interval is 1.0 and the rates range between 0.03 (extremely low) and 3.0 (extremely high), the rates should all be inflated by about 28 - 30%. A hit rate of 0.5 (which yields terrible overall sampling probabilities) produces a bias of about 38%; one of 2.0 (which is downright starry-eyed) produces a bias of 9%. None of this is all so terrible considering that (1) it's hard to imagine there's no clumping of turnover at interval boundaries at all, and (2) a uniform offset will simply change the scaling of the y-axis in an extinction rate plot, which will otherwise look exactly the same.

An entirely unrelated problem is the unfortunate fact that $\ln(p_s)$ is a negative number, which means that in some cases the rates can be negative (!). This is nonsense. What it points to is cruddy data, because it can only happen when binomial error pushes either p_s or N_{3t}/N_{2t} way down. Alternatively, it could be that you are using a global p_s value and sampling in the relevant bin is substantially better than on average (either because you didn't standardize for sampling or standardization didn't work well). So, if a data set produces a lot of negative rates, it's simply a bad idea to compute them no matter what equations you use.

Comparison with Foote's equations.—An interesting property of three-timer rates is that they are consistently higher than Foote face-value rates. For the usual Phanerozoic marine data set the ratio is often around 3:2 (Fig. 5).

I'm not completely sure where this ratio comes from, but it might relate to the fact that the Foote rates include counts of rare, long-ranging taxa and three-timer rates don't. Such taxa qualify as bottom-and-top crossers (N_{bt}) but don't often qualify as first or last appearing in the focal bin (N_t or N_b). Why? Simply because they are rare, and you have to be sampled within the focal bin to be in one of those two categories. In other words, rare, long-ranging can add a lot to both the numerator and denominator of a Foote equation such as $\ln([N_{bt} + N_b]/N_{bt})$, and the more you add the lower the rate falls. If this effect is real it would seem to be a bias, not a signal.

It's probably important that the discrepancy between the rate equations is smaller near the ends of the time series. I doubt that this pattern is related to the Pull of the Recent because three-timer rates can't suffer from it (by definition) and because it can't exist in the early Paleozoic (by definition), where we also see the convergence. More likely, Foote's rates get better toward the ends because censoring (running out of data) makes it harder to sample those nefarious long-ranging but rarely appearing taxa.

Meanwhile, the gap between the rates is largest immediately after the Permo-Triassic and Triassic-Jurassic mass extinctions, where Foote's estimates seem particularly low. This pattern might stem from the Signor-Lipps effect (Signor and Lipps 1982), which is related to censorship. The effect does nothing to counts of truly surviving taxa (N_{bt}), but counts of truly last-appearing taxa (N_b) right before an extinction are systematically too low because these taxa of course can't be sampled again. This problem will bias $\ln([N_{bt} + N_b]/N_{bt})$ downwards as long as undersampling doesn't cause too many taxa to leak from N_{bt} to N_b. Again, I'm only guessing with all of these arguments: as the cliché goes, more study is required.

Three-timer turnover percentages.—Now assume that turnover is not continuous (and therefore exponential), but concentrated in pulses (as suggested by Foote 2005, 2007). A continuous process is defined by a rate sensu stricto, but a pulsed process is a more like a weekly lottery: you can think of the outcome not only

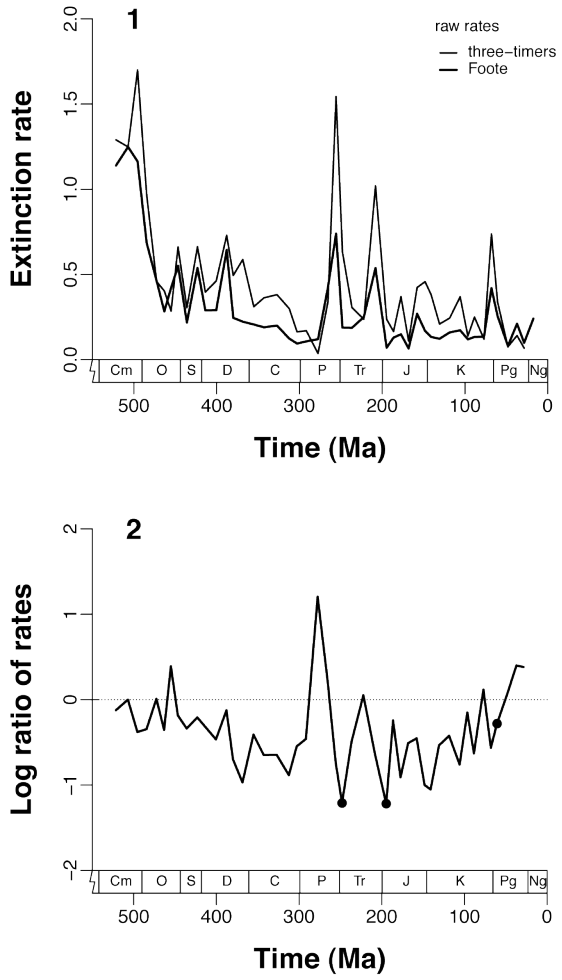

FIGURE *5–1*, *2*. Comparison between alternative instantaneous extinction rate equations. Raw, unstandardized taxon counts taken from Alroy (2008) are used. *1*, Data generated by each method. Thin line = rates generated by the three-timer equation (Alroy 2008); thick line = rates generated by Foote's equation (Foote 1999, 2000). *2*, Log of the ratio between Foote rates and three-timer rates. Black circles = Big Three mass extinctions as defined by Alroy (2008).

in terms of probabilities but in terms of simple fractions or percentages. In these terms, the survival probability is nothing more than the count of taxa continuing past the bin's upper boundary divided by the count of those already present at the bin's base. A ratio of three timers to two timers will give you this probability without presenting any danger of edge effects.

However, this figure is still biased because (as we know) three timers need to be sampled one more time than two timers. Failure to sample in the third bin makes the survival chance seem too small:

$$p_{\text{survive}}\, p_s = N_{3t,i}/N_{2t,i} \tag{19}$$

Therefore, to fix the survival chance you need to do this:

$$p_{\text{survive}} = N_{3t,i}/(N_{2t,i}\, p_s) \tag{20}$$

The death chance p_e is one minus this fraction:

$$p_e = 1 - N_{3t,i}/(N_{2t,i}\, p_s) \tag{21}$$

Likewise, dividing the three timers by the ending two timers gives you the fraction that did not originate in the bin. Thus, the birth chance p_o is one minus this ratio, not forgetting the sampling correction:

$$p_o = 1 - N_{3t,i}/(N_{2t,i+1}\, p_s) \tag{22}$$

Two-timer turnover rates.—The one biggest problem with three-timer rates is random sampling error, which as mentioned can generate negative rates. N_{3t} tends to be a small fraction of the taxa sampled in any particular bin, because to be a three timer you need to (1) survive three intervals and (2) be sampled in three consecutive intervals. This is not very likely in most cases. So, random error tends to be high.

Let's therefore try to squeeze a rate out of N_{2t} and the simple sampled-in-bin count N_s. Again, we want to compute a log ratio of diversity at the start of a bin to the diversity of the same taxa at the end.

The problem is that N_s is not a boundary estimate: it's a function of both standing diversity at the start (N_0) *and* the number of new species that pile up during the bin. Fortunately, Raup (1985) gave an equation that tells us what N_s should be if we know the starting diversity level N_0:

$$N_s = N_0\, (\lambda - \mu \exp(\mu - \lambda))/(\lambda - \mu) \tag{23}$$

Actually, we know N_s, not N_0, but we can rearrange the expression to get N_0:

$$N_0 = N_s/((\lambda - \mu \exp(\mu - \lambda))/(\lambda - \mu)) \quad (24)$$

In theory, now we know μ:

$$\mu = \ln(N_{2t}/N_0) + \ln(p_s) \quad (25)$$

We need the p_s term for the same reason as before: N_{2t} is always too low because some taxa are not sampled in the second bin, but should have been.

But wait a minute. We just used μ (two equations up) to get μ (one equation up). The solution (Alroy 2008, 2009) is to compute N_0 by plugging some arbitrary value of λ into eqn. 24 (say, 0.3), compute μ with eqn. 25, and then go back and do it over again many times (i.e., a recursive computation).

This seemingly counter-intuitive approach turns out to work reasonably well with the Phanerozoic marine data, yielding basically the same numbers that you get with the simpler three-timer equations (Alroy 2008). The advantage is that you can use it with a medium-sized data set and not get too much noise, whereas the three-timer method only works with truly large data sets.

Capture-recapture rate estimates.—Capture-recapture methods for estimating turnover rates have a long history in paleontology (e.g., Nichols and Pollock 1983; Connolly and Miller 2001) and are detailed elsewhere in this volume. They are designed to simultaneously extract diversity, turnover, and sampling intensity estimates from time series data by looking at presences ("captures") and absences of taxa. Foote's equations and traditional diversity curve counting methods are obviously very different because they work with simple age ranges instead. However, three-timer rates and SQS do also involve presence-absence data, and some similar ideas do crop up in the math.

The key difference is that capture-recapture methods address the sampling problem by trying to estimate turnover rates and raw sampling probabilities all at once using the same basic information. Doing so means that error in the sampling estimates biases the turnover rate estimates and vice versa. By contrast, my advice is to clean up the data with sampling standardization first and then compute diversity and turnover rate estimates using different methods. These estimates are independent because sampling standardization algorithms are not influenced by turnover rate estimates: it's a one-way street. Also, because three-timer rate estimates work with short time spans they can be reasonably accurate even when they are not based on standardized taxon counts (see below).

Historically, use of capture-recapture methods is understandable because the ecologists who developed them (e.g., Cormack 1964) employed consistent sampling techniques and uniform amounts of sampling. So, uneven sampling wasn't an issue and incomplete sampling was thought to be a simple nuisance. Because fossil data have such weird sampling properties and because uniform and fair sampling turn out not to be the same, at this stage is remains to be seen whether capture-recapture methods are a useful complement to alternatives such as SQS.

Are three-timer rates really "unbiased"?—Three-timer methods do dispense with a series of predictable, potentially serious difficulties: temporal offsetting of rates caused by the Signor-Lipps, Pull of the Recent, and simple edge effects; inflation of rates due to simple resampling failure; conflation of rates and percentages; and lack of connection to specific evolutionary models. Three-timer methods also make weak assumptions and can be computed easily from sampling-standardized data.

However, no estimate of anything in the fossil record is 100% unbiased, and three-timer rates are no exception. For one thing, there's the continous turnover rate effect problem (although I don't think it's serious). Additionally, three-timer rates do not control for medium-term variation in the effective sampling universe (e.g., thanks to the sampling in consecutive time intervals of unusual environments, geographic regions, or preservational modes, or to a gradual increase in the quality of sampling). However, all the other methods have the same problem, and three-timer rates (like Foote rates) do eliminate sampling universe problems if extraordinary windows of opportunity do not persist beyond single intervals.

Another concern is the potential for covariation of turnover rates and sampling probabilities. If rarely sampled taxa have high rates, then the overall average rate may not be very meaningful. Again, all methods have this problem. More importantly, both the Foote and three-timer equations yield consistent estimates regardless of whether they are computed from raw or sampling-standardized data (Fig. 6).

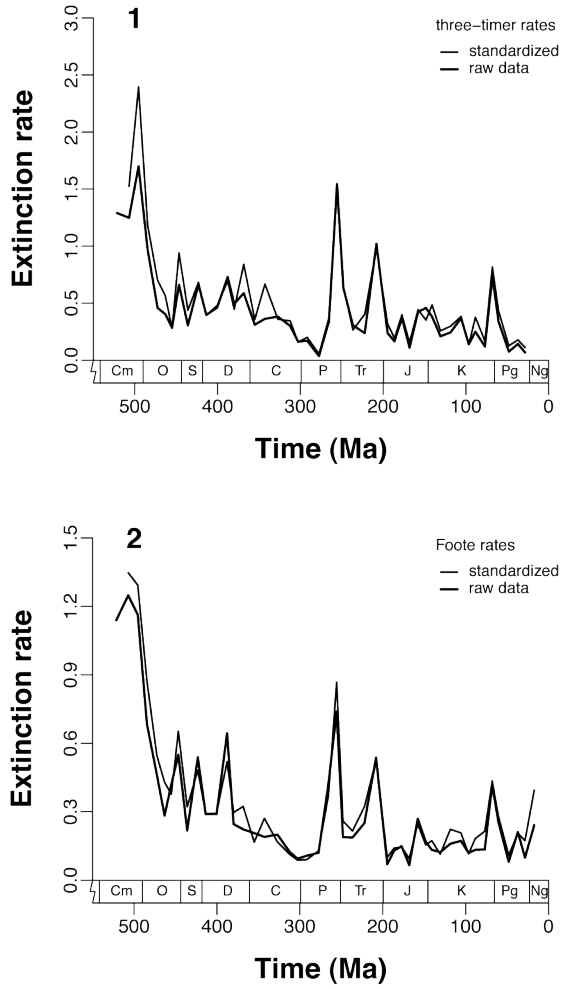

FIGURE *6–1*, *2*.—Consistency of estimates generated by instantaneous extinction rate equations. Thin lines = sampling-standardized data drawn from Alroy (2008); thick lines = raw data draw from the same source. *1*, Rates generated by the three-timer equation (Alroy 2008). *2*, Rates generated by Foote's equation (Foote 1999, 2000).

A likely explanation for the lack of large offsets in either direction is the fact that (1) most of the relevant information is drawn only from a very local part of the time series, and (2) sample size biases tend to be similar in adjacent time intervals. The important point, however, is that raw data sets include many more rare taxa, so they should imply systematically lower rates. Because they don't, any hypothetical covariance between rarity and rates of turnover would appear to have very little effect. And one other thing: because there are many more sampling hits per taxon in the raw data set but the rates are almost the same, little or no bias appears to have been introduced by continuous turnover (see the fine print above).

Capture-recapture methods have some of the advantages of three-timer equations, and Foote's basic range-based rates also escape from many biases. However, both methods do still suffer from the various flavors of edge effects intrinsic to age ranges, and it is not yet clear whether these effects can be removed without introducing further problems. Despite that fact, many researchers still continue to use simple age range data. Foote's equations are the best option if they wish to do so. The time may soon come, however, when neither diversity nor turnover will be quantified by anyone using the simple lists of first and last appearances that constitute analytical paleobiology's classic data sets.

ACKNOWLEDGMENTS

My work in this area was inspired by D. Raup and J. Sepkoski. I thank M. Foote and P. Roopnarine for their insightful reviews and members of the Paleobiology Database's working groups for many helpful methodological discussions. Participants in the Database's 2005 through 2010 analytical paleobiology workshops are in no way responsible for this paper's opacity and lack of coherence. This research was funded privately while the author was a Center Associate at the National Center for Ecological Synthesis and Analysis, and this is Paleobiology Database official publication number 123.

REFERENCES

ALROY, J. 1996. Constant extinction, constrained diversification, and uncoordinated stasis in North American mammals. Palaeogeography, Palaeoclimatology, Palaeoecology, 127:285-311.

ALROY, J. 1998. Equilibrial diversity dynamics in North American mammals. p. 232-287 In M. L. McKinney and J. A. Drake (eds.), Biodiversity Dynamics: Turnover of Populations, Taxa, and Communities. Columbia University Press, New York.

ALROY, J. 1999. The fossil record of North American mammals: evidence for a Paleocene evolutionary radiation. Systematic Biology, 48:107-118.

ALROY, J. 2000a. Successive approximations of diversity curves: ten more years in the library. Geology, 28:1023-1026.

ALROY, J. 2000b. New methods for quantifying macroevolutionary patterns and processes. Paleobiology, 26:707-733.

ALROY, J. 2002. How many named species are valid? Proceedings of the National Academy of Sciences, USA, 99:3706-3711.

ALROY, J. 2008. Dynamics of origination and extinction in the marine fossil record. Proceedings of the National Academy of Sciences, USA, 105:11536-11542.

ALROY, J. 2009. Speciation and extinction in the fossil record of North American mammals. p. 301-323 In R. Butlin, J. Bridle, and D. Schluter (eds.), Speciation and Patterns of Diversity. Cambridge University Press, Cambridge, 346 p.

ALROY, J. 2010. The shifting balance of diversity among major marine animal groups. Science, 329:1191-1194.

ALROY, J. In press. Geographic, environmental, and intrinsic biotic controls on Phanerozoic marine diversification. Palaeontology.

ALROY, J. ET AL. 2001. Effects of sampling standardization on estimates of Phanerozoic marine diversification. Proceedings of the National Academy of Sciences, USA, 98:6261-6266.

ALROY, J. ET AL. 2008. Phanerozoic trends in the global diversity of marine invertebrates. Science, 321:97-100.

BAMBACH, R. K. 1977. Species richness in marine benthic habitats through the Phanerozoic. Paleobiology, 3:152-167.

BAMBACH, R. K. 1999. Energetics in the global marine fauna: a connection between terrestrial diversification and change in the marine biosphere. Geobios, 32:131–144.

BENTON, M. J. 1995. Diversification and extinction in the history of life. Science, 268:52-58.

BENTON, M. J. 2008. How to find a dinosaur, and the role of synonymy in biodiversity studies. Paleobiology, 34:516-533.

BERGER, W. H., AND F. L. PARKER. 1970. Diversity of planktonic Foraminifera in deep-sea sediments. Science, 168, 1345-1347.

BUSH, A. M., AND R. K. BAMBACH. 2004. Did alpha diversity increase during the Phanerozoic? Lifting the veils of taphonomic, latitudinal, and environmental biases. Journal of Geology, 112:625–642.

BUSH, A. M., M. J. MARKEY, AND C. R. MARSHALL. 2004. Removing bias from diversity curves: the effects of spatially organized biodiversity on sampling standardization. Paleobiology, 30:666-686.

CHIARUCCI, A., G. BACARO, D. ROCCHINI, AND L. FATTORINI. 2008. Discovering and rediscovering the sample-based rarefaction formula in the ecological literature. Community Ecology, 9, 121-123.

COLWELL, R. K., AND J. A. CODDINGTON. 1994. Estimating terrestrial biodiversity through extrapolation. Philosophical Transactions of the Royal Society (Series B), 345:101-118.

CONNOLLY, S. R., AND A. I. MILLER. 2001. Joint estimation of sampling and turnover rates from databases: capture-mark-recapture methods revisited. Paleobiology, 27:751-767.

CORMACK, R. M. 1964. Estimates of survival from the sighting of marked animals. Biometrika, 51:429-438.

DARWIN, C. 1859. On the Origin of Species. John Murray, London, 502 p.

FOOTE, M. 1988. Survivorship analysis of Cambrian and Ordovician trilobites. Paleobiology, 14:258-271.

FOOTE, M. 1994. Temporal variation in extinction risk and temporal scaling of extinction metrics. Paleobiology, 20:424-444.

FOOTE, M. 1999. Morphological diversity in the evolutionary radiation of Paleozoic and post-Paleozoic crinoids. Paleobiology, 25(suppl.):1-115.

FOOTE, M. 2000. Origination and extinction components of taxonomic diversity: generaly problems. p. 74-102 In D. H. Erwin and S. L. Wing (eds.), Deep Time: Paleobiology's Perspective. Paleobiology, 26(suppl.).

FOOTE, M. 2001. Inferring temporal patterns of preservation, origination, and extinction from taxonomic survivorship analysis. Paleobiology, 27:602-630.

FOOTE, M. 2005. Pulsed origination and extinction in the marine realm. Paleobiology, 31:6-20.

FOOTE, M. 2007. Extinction and quiescence in marine animal genera. Paleobiology, 33:262-273.

FOOTE, M., AND D. M. RAUP. 1996. Fossil preservation and the stratigraphic ranges of taxa. Paleobiology, 22:121-140.

GILINSKY, N. J., AND I. J. GOOD. 1991. Probabilities of origination, persistence, and extinction of families of marine invertebrate life. Paleobiology, 17:145-166.

GOOD, I. J. 1953. The population frequencies of species and the estimation of population. Biometrika, 40:237-264.

GOTELLI, N. J., AND R. K. COLWELL. 2001. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species diversity. Ecology Letters, 4:379-391.

GOULD, S. J., D. M. RAUP, J. J. SEPKOSKI, Jr., T. J. M. SCHOPF, AND D. S. SIMBERLOFF. 1977. The shape of evolution: a comparison of real and random clades. Paleobiology, 3:23-40.

HARPER, C. W., Jr. 1975. Standing diversity of fossil groups in successive intervals of geologic time: a new measure. Journal of Paleontology, 49:752-757.

HENDY, A. J. W. 2009. The influence of lithification on Cenozoic marine biodiversity trends. Paleobiology, 35:51-62.

KOWALEWSKI, M., W. KIESSLING, M. ABERHAN, F. T. FÜRSICH, D. SCARPONI, S. L. BARBOUR WOOD, AND A. P. HOFFMEISTER. 2006. Ecological, taxonomic, and taphonomic components of the post-Paleozoic increase in sample-level species diversity of marine benthos. Paleobiology, 32:533-561.

KURTÉN, B. 1960. Chronology and faunal evolution of the earlier European glaciations. Commentationes Biologicae, Societas Scientarium Fennici, 21:40-62.

LYELL, C. 1830. Principles of Geology. John Murray, London.

MACARTHUR, R. H., AND E. O. WILSON. 1967. The Theory of Island Biogeography. Princeton University Press, New Jersey, 203 p.

MARSHALL, C. R., AND P. D. WARD. 1996. Sudden and gradual molluscan extinctions in the latest Cretaceous of Western European Tethys. Science, 274:1360-1363.

MAY, R. M. 1975. Patterns of species abundance and diversity. p. 81-120 In M. L.. Cody and J. E. Diamond (eds.), Ecology and Evolution of Communities. Belknap Press of Harvard University Press, Cambridge, Massachusetts, 545 p.

MILLER, A. I. 2000. Conversations about Phanerozoic diversity. p. 53-73 In D. H. Erwin and S. L. Wing (eds.), Deep Time: Paleobiology's Perspective. Paleobiology, 26(suppl.).

MILLER, A. I. AND M. FOOTE. 1996. Calibrating the Ordovician radiation of marine life: implications for Phanerozoic diversity trends. Paleobiology, 22:304-309.

NEWELL, N. D. 1952. Periodicity in invertebrate evolution. Journal of Paleontology, 26:371-385.

NEWELL, N. D. 1959. Adequacy of the fossil record. Journal of Paleontology, 33:488-499.

NICHOLS, J. D., AND K. H. POLLOCK. 1983. Estimating taxonomic diversity, extinction rates, and speciation rates from fossil data using capture-recapture models. Paleobiology, 9:150-163.

NIKLAS, K. J., B. H. TIFFNEY, AND A. H. KNOLL. 1983. Patterns in vascular plant diversification. Nature, 303:614-616.

PAUL, C. R. C. 1982. The adequacy of the fossil record. p. 75-117 In K. A. Joysey and A. E. Friday (eds.), Problems of Phylogenetic Reconstruction. Academic Press, New York, 442 p.

PHILLIPS, J. 1860. Life on Earth: Its Origin and Succession. Macmillan, London, 224 p.

POCOCK, M. J. O., A. C. FRANTZ, D. P., COWAN, P. C. L. WHITE, AND J. B. SEARLE. 2004. Tapering bias inherent in minimum number alive (MNA) population indices. Journal of Mammalogy, 85: 959-962.

PRESTON, F. W. 1948. The commonness, and rarity, of species. Ecology, 29:254-283.

QUENTAL, T. B., AND C. R. MARSHALL. 2009. Extinction during evolutionary radiations: reconciling the fossil record with molecular phylogenies. Evolution, 63:3158-3167.

RAUP, D. M. 1972. Taxonomic diversity during the Phanerozoic. Science, 177:1065-1071.

RAUP, D. M. 1975. Taxonomic diversity estimation using rarefaction. Paleobiology, 1:333-342.

RAUP, D. M. 1976. Species diversity in the Phanerozoic: an interpretation. Paleobiology, 2:289-279.

RAUP, D. M. 1978. Cohort analysis of generic survivorship. Paleobiology, 4:1-15.

RAUP, D. M. 1979. Biases in the fossil record of species and genera. Bulletin of the Carnegie Museum of Natural History, 13:85-91.

RAUP, D. M. 1985. Mathematical models of cladogenesis. Paleobiology, 11:42-52.

RAUP, D. M. 1991. A kill curve for Phanerozoic marine species. Paleobiology, 17:37-48.

RAUP, D. M., S. J. GOULD, T. J. M. SCHOPF, AND D. SIMBERLOFF. 1973. Stochastic models of phylogeny and the effect of diversity. Journal of Geology, 81:525-542.

RAUP, D. M., AND J. J. SEPKOSKI, JR. 1982. Mass extinctions in the marine fossil record. Science, 215:1501-1503.

RAUP, D. M., AND J. J. SEPKOSKI, JR. 1984. Periodicity of extinctions in the geologic past. Proceedings of the National Academy of Sciences, USA, 81:801-805.

RAUP, D. M., AND S. M. STANLEY. 1971. Principles of Paleontology. W. H. Freeman, San Francisco, 388 p.

RUDWICK, M. J. S. 1998. George Cuvier, Fossil Bones, and Geological Catastrophes: New Translations and Interpretations of the Primary Texts. University of Chicago Press, Chicago, 318 p.

SANDERS, H. L. 1968. Marine benthic diversity: a comparative study. American Naturalist, 102:243-282.

SEPKOSKI, D. 2005. Stephen Jay Gould, Jack Sepkoski, and the 'quantitative revolution' in American paleobiology. Journal of the History of Biology, 38:209-237.

SEPKOSKI, J. J., JR. 1975. Stratigraphic biases in the analysis of taxonomic survivorship. Paleobiology, 1:343-355.

SEPKOSKI, J. J., JR. 1978. A kinetic model of Phanerozoic taxonomic diversity. I. Analysis of marine orders. Paleobiology, 4:223-251.

SEPKOSKI, J. J., JR. 1979. A kinetic model of Phanerozoic taxonomic diversity. II. Early Phanerozoic families and multiple equilibria. Paleobiology, 5:222-251.

SEPKOSKI, J. J., JR. 1982. A compendium of fossil marine families. Milwaukee Public Museum Contributions in Biology and Geology, 51:1-125.

SEPKOSKI, J. J., JR. 1984. A kinetic model of Phanerozoic taxonomic diversity. III. Post-Paleozoic families and mass extinctions. Paleobiology, 10:246-267.

SEPKOSKI, J. J., JR. 1990. The taxonomic structure of periodic extinction. Geological Society of America Special Paper, 247:33-44.

SEPKOSKI, J. J., JR. 1997. Biodiversity: past, present, and future. Journal of Paleontology, 71:533-539.

SEPKOSKI, J. J., JR. 2002. A compendium of fossil marine animal genera. Bulletins of American Paleontology, 363:1-560.

SEPKOSKI, J. J., JR., R. K. BAMBACH, D. M. RAUP, AND J. W. VALENTINE. 1981. Phanerozoic marine diversity and the fossil record. Nature, 293:435-437.

SESSA, J. A., M. E. PATZKOWSKY, AND T. J. BRALOWER. 2009. The impact of lithification on the diversity, size distribution, and recovery dynamics of marine invertebrate assemblages. Geology, 37:115-118.

SHINOZAKI, K. 1963. Note on the species area curve. Proceedings of the 10th Annual Meeting of the Ecological Society of Japan, 5.

SIGNOR, P. W., III, AND J. H. LIPPS. 1982. Sampling bias, gradual extinction patterns, and catastrophes in the fossil record. Geological Society of America Special Publication, 190:291-296.

SIMPSON, G. G. 1944. Tempo and Mode in Evolution. Columbia University Press, New York, 237 p.

SIMPSON, G. G. 1949. The Meaning of Evolution. Yale University Press, New Haven, Connecticut, 364 p.

SIMPSON, G. G. 1952. Periodicity in vertebrate evolution. Journal of Paleontology, 26:359-370.

SIMPSON, G. G. 1960. The history of life. p. 117-180 In S. Tax (ed.), Evolution After Darwin. Volume 1: The Evolution of Life. University of Chicago Press, Chicago, 629 p.

SMITH, E. P., P. M. STEWART, AND J. CAIRNS, JR. 1985. Similarities between rarefaction methods. Hydrobiologia, 120:167-170.

STANLEY, S. M. 1973. Effects of competition on rates of evolution, with special reference to bivalve molusks and mammals. Systematic Zoology, 22:486-506.

STANLEY, S. M. 1975. A theory of evolution above the species level. Proceedings of the National Academy of Sciences, USA, 72:646-650.

TARVER, J. E., S. J. BRADDY, AND M. J. BENTON. 2007. The effects of sampling bias on Paleozoic faunas and implications for macroevolutionary studies. Palaeontology, 50:177-184.

TIPPER, J. C. 1979. Rarefaction and rarefiction: the use and abuse of a method in paleoecology. Paleobiology, 5:423-434.

VALENTINE, J. W. 1970. How many marine invertebrate fossil species? A new approximation. Journal of Paleontology, 44:410-415.

VAN VALEN, L. 1973. A new evolutionary law. Evolutionary Theory, 1:1-30.

WEBB, S. D. 1969. Extinction-origination equilibria in late Cenozoic land mammals of North America. Evolution, 23:688-702.