Likelihood tests of hypothesized durations: determining and accommodating biasing factors

Peter J. Wagner

Abstract.—Paleobiologists frequently hypothesize that a taxon's duration (i.e., the true span from origination to extinction) exceeds its stratigraphic range (i.e., the span from first appearance to last appearance in the fossil record). One can test hypothesized duration by assessing the plausibility of the implicitly hypothesized gaps between origination and first appearance and/or between last appearance and extinction. Several tests assess the probability of not finding a taxon over some stratigraphic gap. Because the likelihood of a hypothesis given data reflects the probability of the data given that hypothesis, these probabilities also give the likelihood of a hypothesized duration. However, many probability/likelihood tests require simplifying assumptions about unknown sampling parameters such as the consistency of sampling over time, sampling intensities for unknown ancestors, and actual sampling intensities themselves.

This paper examines the effects of sampling parameters on probability/likelihood tests and presents methods for testing hypotheses about these unknowns while testing hypotheses about true durations. Two data sets are used here as examples. One analysis tests the origination times among Paleozoic gastropods implied by phylogenetic inferences. The other analysis tests the extinction times among Maastrichtian ammonites implied by different numbers of extinction events. In both cases, hypotheses positing many gaps in the fossil record become more likely after accommodating uncertainty about sampling. However, the increased likelihoods are insufficient to prevent these hypotheses from being rejected in favor of hypotheses positing fewer gaps. In both cases, the conclusions are identical to those derived by simple methods using simple models for unknown sampling parameters. Although numerous factors can exaggerate the implausibility of gaps, making these factors parts of testable hypotheses is possible. Thus, excessive assumptions about sampling parameters need not hinder empirical testing of hypothesized durations.

Peter J. Wagner. 1400 South Lake Shore Drive, Department of Geology, Field Museum of Natural History, Chicago, Illinois 60605. E-mail: pwagner@fmnh.org

Accepted: 22 November 1999

Introduction

Hypotheses about the pace of mass extinctions, divergence times between sister taxa, and paleocommunity stability all make predictions about originations and/or extinctions. In doing so, these hypotheses implicitly posit durations (i.e., the spans between true originations and true extinctions). Frequently, hypothesized durations will exceed observed stratigraphic ranges (i.e., intervals from first appearance to last appearance) and thus posit a sampling gap between either the hypothesized origination and first appearance, or the last appearance and the hypothesized extinction. Methods that can test hypothesized durations clearly are important to many paleobiological research programs. Hypothesized durations can make probabilistic predictions about stratigraphic ranges (Paul 1982, 1985; Springer and Lilje 1988; Strauss and Sadler 1989). Numerous studies therefore use probability or likelihood to test sampling gaps implied by hypothesized times of extinction (e.g., Springer 1990; Marshall 1995a; Marshall and Ward 1996) and origination (Marshall 1990, 1995b; Cheetham and Jackson 1995; Wagner 1995; Huelsenbeck and Rannala 1997; Gingerich and Uhen 1998; Foote et al. 1999). However, hypothesized durations alone do not make predictions about stratigraphic data unless one also considers other "unknowns." These unknowns include sampling intensity (i.e., the probability of sampling a taxon over some stratigraphic interval), temporal consistency of sampling intensity, and the sampling intensities of unknown ancestral species. This has led some workers to question the soundness of conclusions drawn from probability and likelihood tests because the tests assume conditions for too many unknown parameters (e.g., Smith 1994; Rieppel 1997). Given these justified criticisms, it is especially important
to develop probability and likelihood tests that can identify and test unknowns other than hypothesized durations.

Methods for testing hypotheses about actual sampling intensities and variation in those intensities have long been used to test hypotheses ranging from declining product failure and accidents in industry to changing performances by professional athletes. Previous attempts to address fossil sampling parameters (e.g., Marshall 1990, 1994, 1997) offer tests of sampling consistency and methods for calculating confidence intervals on taxon durations with nonrandom distributions of fossil finds. However, those works do not present a protocol for calculating the likelihoods of hypothesized gaps once such complicating factors are identified. This paper presents likelihood approaches that explicitly test hypotheses about these unknowns while testing hypotheses about durations at the same time. I then apply these methods to two previously presented studies and examine whether accommodating biases affects the conclusions drawn from the data sets.

Unknowns such as sampling intensity, sampling consistency, and sampling intensities for hypothesized ancestors are models (i.e., assumptions or premises) in many analyses. A primary goal of this paper is to separate models (i.e., “that part of the description which is not at present in question, and may be regarded as given...” [Edwards 1992: p. 3]) from hypotheses (“parameters or entities being in question, and the subject of investigation” [Edwards 1992: pp. 3–4]). The model of one analysis can be the hypothesis of another: for example, Pearson correlation tests use a normal distribution as a model when testing for independent distributions whereas Kolmogorov–Smirnov analyses can test and reject the hypothesis that a variable is normally distributed. This paper focuses on testing hypotheses about these unknowns rather than assuming particular conditions for them. The paper then will develop tests of multiparameter hypotheses in which a hypothesized duration is one of the unknown parameters being tested.

**Basic Likelihood Tests of Hypothesized Durations**

The hypothesis that an organism is still extant at a certain point in time is one that can be proved but never disproved. (The same also is true for a hypothesis that a taxon existed for some duration prior to its first appearance in the fossil record.) Thus, the hypothesis that a taxon is extant cannot be falsified sensu stricto. However, if there are many observations in which a taxon could have been sampled but was not, then the hypothesis that it still exists becomes increasingly implausible. At some point, the plausibility will become so low that the hypothesis should no longer be seriously entertained. Consider an example where a taxon is observed over 11 time units (Fig. 1). These units could be stratigraphic intervals, fossiliferous horizons arranged in stratigraphic order, or ecological surveys conducted over some period. Given the number of samples in which it is found and the number of samples within its range in which it is not found, we can test the hypothesis that the taxon still is extant in time unit 16.

Likelihood tests of duration hypotheses using range data and distributions of observations have been developed both for paleobiology and conservation biology (see, e.g., Solow 1993a; Burgman et al. 1995). Estimating the likelihood that a taxon is already (or still) present requires a stratigraphic range, \( R \) (after Marshall 1990), and an observed number of finds, \( n \) (after Huelsenbeck and Rannala 1997 and Solow and Smith 1997; \( = H \) of Marshall 1990). Note that “range” here is the range of a taxon through a series of correlated strata, not simply a local range. Because ranges are inferences for fossil taxa, observed ranges are not truly data. However, I will assume here that accurate correlations for inferring ranges are available. Huelsenbeck and Rannala (1997: eq. 2) estimated the likelihood of a hypothesized duration \( D \) given observed range \( R \) and finds \( n \) as

\[
L[D \mid n, R] = k \frac{(R - 1)^{n-2} \cdot \theta^{n-2} \cdot e^{-n(D-1)}}{(n - 2)!}
\]

when \( n \geq 2 \)  

where \( k \) (here and elsewhere) is a constant that scales the highest probability of the observed data to 1.0. \( \theta \) is the sampling intensity (preservation rate) (see discussion below). \( \theta \) here equals Huelsenbeck and Rannala’s \( \lambda \) and
FIGURE 1. A hypothetical situation where a species has not been sampled over some period of time and might be extinct. The circles represent samples that could have revealed the species. These might be ecologic surveys for Recent taxa, or stratigraphic units in which taphonomically and environmentally similar taxa are sampled (see Bottjer and Jablonski 1988). The hypothesis that the taxon still is extant can never be disproved. However, the likelihood of this hypothesis decreases with the accumulation of samples after the last known appearance (i.e., $T_{11}$) that lack the species.

Because the actual sampling intensity is unknown, sampling intensity is a model here.
In some cases, a hypothesized duration predicts one gap, either before the first known appearance or after the last known appearance. The likelihood of this duration is proportional to the probability of one find in the first (last) appearance followed by failure to sample over the gap. This can be calculated as

$$L[D|n, R] = k \frac{(G + 1) \theta \cdot e^{-\xi(G+1)}}{G + 1}$$

The gap, $G$, is simply the difference between the hypothesized origination (extinction) and the first (last) appearance (i.e., $D - R$, with the
hypothesized gap only one end of the range). The numerator gives the Poisson probability of a single find given some expected number of finds (i.e., \( \theta \cdot [G + 1] \)). However, this is the probability of a single find anywhere over \( G + 1 \) hypothesized sample opportunities (i.e., the hypothesized gap plus the first or last appearance). Here we know that the lone find is in the last possible horizon if testing a hypothesis of extinction (e.g., time unit 11 in Fig. 1) or in the first if testing a hypothesis of origination. Thus, the actual probability is the Poisson probability divided by \( G + 1 \). Note that our hypothesized duration includes only one extension, either for extinction or origination, and really is a hypothesis about the gap, \( G \), between the hypothesized extinction (or origination) and the last (or first) appearance.

Equation (2) is directly proportional to equation (1), differing by a constant:

\[
\frac{(R - 1)^{-3} \cdot \theta^{-3} \cdot e^{-\theta(R-2)}}{(n - 2)!}.
\]

Because one scales likelihoods so that the highest likelihood is 1.0 (Edwards 1992), equations (1) and (2a) ultimately give the same likelihoods. Thus, both equations test identical hypotheses.

Equations (1) and (2a) test the probability that a taxon is still (already) present by calculating the probability of \( X \) finds over some interval given the total number of finds over a more restricted interval. This is appropriate for testing the hypothesis that a particular morphotype was still or already present. However, there are circumstances in which workers are interested in the probability of not sampling an unspecified taxon over some interval. For example, phylogenetic range extensions posit that a species or some unsampled ancestral morphotype originated by a particular point in time. Here we need to test the probability of no finds over a gap spanning some number of sampling opportunities. This is given by several workers (e.g., Foote 1997: p. 297; Huelsenbeck and Rannala 1997: p. 176) simply as

\[
L[D|n = 0] = e^{-\theta G}.
\]  

(3a)

A Poisson distribution is appropriate if one’s sampling allows multiple finds per bin (e.g., \( n = 3 \) over one meter of sediment, with ranges described in meters of sediment, or \( n = 3 \) over 1 million years, with ranges described in millions of years). However, sampling tallies often allow only one find per bin (e.g., simple presence/absence per fossiliferous horizon or formation). In such cases, Alroy (personal communication in Wagner 1995) recommends also expressing \( R, D, \) and \( G \) in discrete numbers of sampling opportunities over an observed range or a hypothesized duration. Figure 1 uses this approach. If discrete sampling units are used, then a binomial distribution is more appropriate than a Poisson distribution because the former gives the probability of 0 to \( X \) finds given \( X \) opportunities whereas the latter gives the probability of 0 to \( \infty \) finds over duration \( X \). When assuming a binomial distribution, equation (2a) becomes

\[
L[D|n, R] = k \cdot \theta \cdot (1 - \theta)^n
\]  

(2b)

and equation (3a) becomes

\[
L[D|n = 0] = k \cdot (1 - \theta)^n
\]  

(3b)

The likelihood of durations and their implied gaps always will be lower when assuming a binomial distribution than when assuming a Poisson distribution. This is desirable if the absence of a gap is the null hypothesis. However, when the null hypothesis is that a gap exists, then assuming a Poisson distribution is conservative even if the binomial distribution is more appropriate.

**Sampling Intensity As a Model and As a Hypothesis**

**Calculating Sampling Intensity (\( \theta \)).**—Although the equations above are described as the likelihoods of a hypothesis given ranges and finds, the equations really determine the likelihood of a hypothesized duration given sampling intensity (\( \theta \)). However, sampling intensity is itself an unknown. A common solution is to use an empirical estimate to model sampling intensity. An intuitive empirical estimate would be the number of finds over the range or number of possible finds:

\[
\theta = \frac{n}{R}.
\]
For example, in Figure 1 we would estimate a sampling intensity of 0.64 finds per sampling opportunity. However, the actual number of sampling opportunities is not the range but the duration. Because the range usually underestimates the duration, \( n/R \) is a biased estimator of sampling intensity (e.g., Marshall 1995a; Solow and Smith 1997). Simple simulations confirm this, indicating that \( n/R \) overestimates sampling intensity appreciably (Fig. 2A), especially when true sampling intensity is low (Fig. 2B). Because overestimating sampling intensity leads likelihood tests to exaggerate the implausibility of hypothesized durations, this clearly is undesirable.

A taxon's range is determined by its first and last finds, which means that we know where two of the finds must occur. However, the span of \( R - 2 \) samples between the first and last appearance is not determined by sampling, but by the taxon's duration. Thus, Foote and Raup (1996) noted that sampling intensity predicts the number of finds between the first and last appearances. If we tally simple presence/absence, then there are \( n - 2 \) finds distributed over that span (with the first and last determining the range and thus outside the span). Therefore, Alroy (1998) estimates sampling intensity as

\[
\theta = \frac{n - 2}{R - 2}.
\]

Simulations show that \( (n - 2)/(R - 2) \) presents a nearly unbiased estimator of sampling intensity (Fig. 2A) and performs far better than \( n/R \) (Fig. 2B).

Calculating sampling intensity as \( (n - 2)/(R - 2) \) not only provides an unbiased estimate of sampling intensity, it also makes likelihood tests more conservative by assuming lower sampling intensities than assumed when using \( n/R \). It also requires that a taxon be known from at least three horizons to calculate gap likelihoods.

Testing the Likelihoods of \( \theta \) and \( D \) Simultaneously. —Equations (1), (2), and (3) all treat sampling intensity as a model. Equation (4) offers an empirical, unbiased estimate that can parameterize that model. A preferable approach, however, is to test hypotheses about sampling intensity. I noted above that sampling intensity predicts the number of finds over the \( R - 2 \) samples between the first and last appearances. This also means that the likelihood of a sampling intensity is proportional to the probability of \( n - 2 \) finds over \( R - 2 \) samples given that sampling intensity (see Fig. 3A). We now can test the likelihood of a hypothesized duration and a hypothesized sampling intensity using the product of the likelihoods of both hypotheses given observed data:

\[
L[D, \theta | n, R] = k \cdot L[D | n, R] \cdot L[\theta | n, R].
\]

The first term, \( L[D | n, R] \), is either equation (2a) or (3a) when assuming a Poisson distribution of finds and misses, or equation (2b) or (3b) when assuming a binomial distribution of finds and misses. The second term is

\[
L[\theta | n, R] = k \cdot \frac{(R - 2)! \theta^{n-2} e^{-\theta (R - 2)}}{(n - 2)!}.
\]

when assuming a Poisson distribution of finds and misses, and

\[
L[\theta | n, R] = k \cdot \theta^{n-2} (1 - \theta)^{(R - n)}.
\]

when assuming a binomial distribution. Under either equation, the most likely sampling intensity is given by equation (4). However, if \( R \) and (especially) \( n \) are low, then much lower sampling intensities also will have high likelihoods.

For any hypothesized gap, the first derivative of the likelihood curve will equal zero at the value of sampling intensity that maximizes \( L[D, \theta | n, R] \) (Edwards 1992: p. 72). The calculus usually is simplest when using the support curve, which plots the natural logarithm of the likelihood equation. If equation (2a) is used to test a hypothesized duration, then the likelihood of a hypothesized gap and sampling intensity is maximized when (see Appendix 1)

\[
\theta = \frac{n - 1}{D - 1}.
\]

Appendices 2–4 give analogs of equation (7) based on equations (2b), (3a), and (3b).

Treating sampling intensity as a hypothesis has appreciable effects, especially when hypothesized gaps are large (Fig. 3B). If one is testing a hypothesis that a particular lineage
Figure 2. The accuracy of two estimates of sampling intensity ($\theta$) in Monte Carlo simulations. The Monte Carlo runs evolve lineages with an extinction rate of 0.01 per time unit. There is one possible find per stratigraphic bin, with the probability of finding a taxon in each possible find equal to $\theta$. Note that a stratigraphic bin might be a fossiliferous horizon, a set amount of sediment (e.g., one meter), or even a whole formation. After extinction, ranges are determined from the first and last samples and sampling intensity is calculated both as $n/R$ and as $(n - 2)/(R - 2)$. A, Average estimates of sampling intensity ($\theta$) at different actual intensities. B, Average deviations from actual $\theta$’s. Based on 1000 simulated ranges with $n \geq 2$ for each true $\theta$. 
was still (already) present at a particular interval (and thus predicting a particular gap), then the increased likelihood of the gap given lower sampling intensity begins to exceed the decreased likelihood of that lower sampling intensity (Fig. 3B). For example, if a taxon is found five times over 20 horizons, then the likelihood of a gap of 100 horizons is orders of magnitude more likely when sampling intensity is treated as a hypothesis rather than a model (\(\ln L[D | n, R, n] = -7.4\) versus \(\ln L[D | R, n] = -16.7\); see Fig. 3B). Nevertheless, a log likelihood of \(-7.4\) still renders the hypothesis highly unlikely given observed sampling. One would reject any hypothesized duration implying a gap of 14 or more sampling
opportunities (using the standard rejection criterion of a difference in support of 2).

**Testing the Likelihoods of \( D \) against All Possible \( \theta \).—**One can determine the likelihood given all possible sampling intensities by integrating equation (5) for \( \theta = 0 \) to \( \theta = 1 \). Although computationally cumbersome, computer programs such as Mathematica* or MathCad* can calculate such integrals effectively. However, applying this approach to the examples used in this paper makes little difference: the relative likelihoods for alternative hypothesized durations are extremely similar to the relative likelihoods found when solving for both duration and sampling intensity. Because sampling intensity is an interesting question itself, I will solve for both hypotheses in the examples below.

**Testing for and Accommodating Variable Sampling Intensities (\( \theta \)).—**The test of hypothesized gaps and sampling intensities given above assumes a single sampling intensity. However, sampling intensity might vary over a taxon’s range. Variable (“nonstationary”) sampling intensity induces underestimates or overestimates of gap likelihoods (Solow 1993b), especially if sampling intensity of the first or last interval differs appreciably from average sampling intensity. Marshall (1990) recommends chi-square or Kolmogorov-Smirnov analyses to test for deviations from consistent sampling intensity. Conservation biologists have addressed the same issue by testing before solving for duration and sampling intensity. Because sampling intensity is an interesting question itself, I will solve for both hypotheses in the examples below.

**Figure 4.** Testing hypotheses that sampling intensities vary over time. The most likely single sampling intensity (\( \theta_i \)) for the entire range is 11/25 = 0.44. The log-likelihood of this given three intervals is \( \ln L[\theta, R_1, R_2, R_3] = -4.44 \). Sampling appears to be denser in intervals 2–3 than in interval 1, which leads to a two-parameter hypothesis: one sampling intensity for interval 1 (\( \theta_1 = 2/8 = 0.25 \)) and another for intervals 2-3 (\( \theta_{2,3} = 9/17 = 0.53 \)). The log-likelihood of this hypothesis is proportional to \( \ln L[\theta, R_1, R_2, R_3] = -1.14 -1.42 -1.25 = -3.81 \). The same approach is used to assess a three-parameter hypothesis of different sampling intensities for each interval. Assessments of significance are discussed in the text.
\[ \theta_3 = \frac{n_{3-1}}{R_{3-1}} = \frac{4}{7} \]

\[ \theta_{2-3} = \frac{n_{2-3-1}}{R_{2-3-1}} = \frac{9}{17} \]

\[ \theta_2 = \frac{n_2}{R_2} = \frac{5}{10} \]

\[ \theta_{1-3} = \frac{n_{1-3-2}}{R_{1-3-2}} = \frac{11}{25} \]

\[ \theta_1 = \frac{n_{1-1}}{R_{1-1}} = \frac{2}{8} \]

\[ \theta_1 = \frac{n_{1-1}}{R_{1-1}} = -\frac{2}{8} \]

- Samples without taxon (observed)
- Samples with taxon (observed)
Table 1. Hypothetical distributions of finds. \( R_i \) gives the number of possible finds for each interval, \( n_i \) gives the number of finds for taxon \( t \) in interval \( i \) and \( \theta_i \) gives the most likely estimate of sampling intensity for taxon \( t \) in interval \( i \). Data and statistics for the entire range (162 horizons) also are given. Note that \( \theta_i \) for intervals I and VI is \( n_i^{-1}/R_i-1 \) because the taxa range through one end of either interval. However, \( \theta_i \) for intervals II–V are \( n_i/R_i \) because the taxa range through both ends of each interval.

<table>
<thead>
<tr>
<th></th>
<th>Entire range</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
<th>VI</th>
</tr>
</thead>
<tbody>
<tr>
<td>( R_i )</td>
<td>162</td>
<td>26</td>
<td>27</td>
<td>28</td>
<td>26</td>
<td>30</td>
<td>25</td>
</tr>
<tr>
<td>( n_{3i} )</td>
<td>70</td>
<td>10</td>
<td>16</td>
<td>10</td>
<td>8</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>( n_{3i} )</td>
<td>66</td>
<td>6</td>
<td>7</td>
<td>20</td>
<td>9</td>
<td>13</td>
<td>11</td>
</tr>
<tr>
<td>( \theta_{3i} )</td>
<td>0.425</td>
<td>0.360</td>
<td>0.593</td>
<td>0.357</td>
<td>0.308</td>
<td>0.333</td>
<td>0.583</td>
</tr>
<tr>
<td>( \theta_{3i} )</td>
<td>0.400</td>
<td>0.200</td>
<td>0.259</td>
<td>0.714</td>
<td>0.346</td>
<td>0.433</td>
<td>0.417</td>
</tr>
</tbody>
</table>

\( R_i = 1 \). A similar argument holds for the final interval (Fig. 4).

The hypothesis that there are \( Y \) sampling intensities over those \( X \) intervals is an alternative to the single-sampling-intensity hypothesis. Given the data in Figure 3, one might hypothesize one sampling intensity for interval 1 and another spanning intervals 2–3, which gives us \( Y = 2 \) and \( X = 3 \). The likelihood of this hypothesis is

\[
L[\theta_1 \ldots \theta_Y | n_1 \ldots n_X, R_1 \ldots R_X] = k\prod_{j=1}^{Y} \prod_{i=1}^{X} L[\theta_i | n_{ij}, R_{ij}] \tag{9}
\]

(see Cox and Lewis 1966: p. 231). Each \( Z_j \) is the number of intervals for which the \( j \)th sampling intensity \( (\theta_j) \) applies. In our example (Fig. 4), \( Z_1 = 1 \) whereas \( Z_2 = 2 \). Each \( n_{ij} \) and \( R_{ij} \) is the number of finds and possible finds for the \( i \)th interval within the \( j \)th sampling intensity set. Thus, \( n_{11} = 3 \) (the finds in interval 1), \( n_{21} = 5 \) (the finds in interval 2) and \( n_{22} = 5 \) (the finds in interval 3). Note that equation (9) reduces to equation (8) if only one sampling intensity is hypothesized.

Equation (9) will be maximized when \( Y = X \), i.e., when there is a separate sampling intensity for each interval. The question now is whether such a complex hypothesis predicts the data significantly better than simpler hypotheses. In our example above, we can test whether a two-parameter hypothesis (e.g., \( \theta_1 \) applying to intervals 1 and \( \theta_3 \) for intervals 2–3) might have nearly as great a likelihood as the three-parameter hypothesis. Likelihood ratio analyses can test whether added predictive power justifies three extra parameters. The test statistic is

\[
\Lambda = 2(\ln L[H_{A+B}|\text{data}] - \ln L[H_A|\text{data}])
\]

where \( H_A \) is the hypothesis with \( A \) parameters and \( H_{A+B} \) is the hypothesis with \( A + B \) (Sokal and Rohlf 1981: p. 695). \( \Lambda \)’s expected distribution is approximately \( \chi^2 \), with degrees of freedom equaling the difference in parameters (i.e., \( B \)), with \( H_A \) considered the null hypothesis because it invokes fewer parameters. Rejecting a low-parameter hypothesis in favor of a high-parameter hypothesis requires a large difference in likelihoods. There are two important points here: (1) hypotheses of heterogeneous sampling usually predict stratigraphic distributions better than do hypotheses of homogenous sampling, but (2) homogenous sampling represents an appropriate null hypothesis because it provides a single explanation for the observed distributions.

Application of this test to Figure 3 shows that we cannot reject a one-parameter hypothesis in favor of more complicated ones despite the apparent variation in sampling over time. This is due largely to the small numbers involved. Consider instead two hypothetical taxa, \( M. mcgwirei \) and \( S. sosai \), that are known from 70 and 66 horizons respectively from a common stratigraphic range (Table 1). Suppose further that this range spans 162 horizons in which either taxon could be found, given occurrences of taphonomically and environmentally similar taxa (Bottjer and Jablonski 1988). Finds and possible finds are binned into six successive units, but exact correlations within those six units are unavai-
Phylogenies offer an alternative way to treat sampling intensity as a hypothesis. The calculations of sampling intensity outlined above are appropriate for testing hypotheses about a single morphotype (i.e., the units for which n and R typically are recorded). This presents little problem for extinction hypotheses, which make no predictions about the extinction of distinguishable morphotypes (but see Archibald 1993). A greater problem is that phylogenetic hypotheses predict the divergence times of lineages but they do not predict the appearance (or disappearance) times of particular morphotypes. Norell (1993) separated phylogenetic range extensions into two types: "ghost lineages" (hypothesized extensions of a sampled lineage) and "ghost taxa" (hypothesized unsampled ancestors of sampled lineages). The likelihood that taxon B arose at the same time as a putative sister taxon A tests hypotheses about ghost lineages, but not necessarily about ghost taxa. This is especially true if sampling intensities vary greatly among close relatives, which decreases our confidence that unsampled ancestors had the same sampling intensities as sampled descendants.

Assuming that unknown ancestors had very low sampling intensities might seem to negate the results of probability tests. However, this alternative has even less basis in theory than does assuming that ancestral sampling intensities equaled descendant sampling intensities. Biological factors affecting preservation potential (i.e., morphologic, ecological, environmental preference, and metapopulational characteristics [see, e.g., McKinney 1986a, 1986b]) are not randomly distributed within phylogenies (Raup and Gould 1974; Felsenstein 1985). It follows logically that if the determinants of sampling intensity are not distributed randomly within phylogeny, then neither should be sampling intensity. Therefore, a hypothesized phylogeny presents a hypothesis about changes in sampling intensity and ancestral sampling intensities just as it does about the factors determining sampling intensity.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>—</td>
<td>0.20</td>
<td>0.24</td>
<td>0.18</td>
<td>0.42</td>
</tr>
<tr>
<td>2</td>
<td>1.68</td>
<td>—</td>
<td>0.28</td>
<td>0.20</td>
<td>0.50</td>
</tr>
<tr>
<td>3</td>
<td>2.83</td>
<td>1.12</td>
<td>—</td>
<td>0.15</td>
<td>0.54</td>
</tr>
<tr>
<td>4</td>
<td>4.91</td>
<td>3.24</td>
<td>2.08</td>
<td>—</td>
<td>0.99</td>
</tr>
<tr>
<td>5</td>
<td>5.01</td>
<td>3.33</td>
<td>2.18</td>
<td>0.10</td>
<td>—</td>
</tr>
</tbody>
</table>

**Table 2.** Likelihood tests of consistent (stationary) sampling intensity (θ) for *M. mcgwirei*. Four hypotheses are contrasted: consistent θ, two θ's (0.390 for intervals I–V and 0.583 for intervals VI), three θ's (0.444 for intervals I–III, 0.321 for intervals IV–V, and 0.583 for interval VI), four θ's (0.360 for interval I, 0.593 for interval II, 0.333 for intervals III–V, and 0.583 for interval VI), and a unique θ for each interval. Lower left gives the Δ statistic (i.e., twice the difference in log-likelihood). The calculations of sampling intensity outlined above are appropriate for testing hypotheses about a single morphotype (i.e., the units for which n and R typically are recorded). This presents little problem for extinction hypotheses, which make no predictions about the extinction of distinguishable morphotypes (but see Archibald 1993). A greater problem is that phylogenetic hypotheses predict the divergence times of lineages but they do not predict the appearance (or disappearance) times of particular morphotypes. Norell (1993) separated phylogenetic range extensions into two types: "ghost lineages" (hypothesized extensions of a sampled lineage) and "ghost taxa" (hypothesized unsampled ancestors of sampled lineages). The likelihood that taxon B arose at the same time as a putative sister taxon A tests hypotheses about ghost lineages, but not necessarily about ghost taxa. This is especially true if sampling intensities vary greatly among close relatives, which decreases our confidence that unsampled ancestors had the same sampling intensities as sampled descendants.

Assuming that unknown ancestors had very low sampling intensities might seem to negate the results of probability tests. However, this alternative has even less basis in theory than does assuming that ancestral sampling intensities equaled descendant sampling intensities. Biological factors affecting preservation potential (i.e., morphologic, ecological, environmental preference, and metapopulational characteristics [see, e.g., McKinney 1986a, 1986b]) are not randomly distributed within phylogenies (Raup and Gould 1974; Felsenstein 1985). It follows logically that if the determinants of sampling intensity are not distributed randomly within phylogeny, then neither should be sampling intensity. Therefore, a hypothesized phylogeny presents a hypothesis about changes in sampling intensity and ancestral sampling intensities just as it does about the factors determining sampling intensity.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>—</td>
<td>0.014</td>
<td>0.007</td>
<td>0.017</td>
<td>0.063</td>
</tr>
<tr>
<td>2</td>
<td>6.081</td>
<td>—</td>
<td>0.048</td>
<td>0.129</td>
<td>0.354</td>
</tr>
<tr>
<td>3</td>
<td>9.992</td>
<td>3.911</td>
<td>—</td>
<td>0.666</td>
<td>0.920</td>
</tr>
<tr>
<td>4</td>
<td>10.178</td>
<td>4.097</td>
<td>0.186</td>
<td>—</td>
<td>0.857</td>
</tr>
<tr>
<td>5</td>
<td>10.485</td>
<td>4.405</td>
<td>0.494</td>
<td>0.308</td>
<td>—</td>
</tr>
</tbody>
</table>

**Table 3.** Likelihood tests of consistent (stationary) θ for *S. sosai*. Four hypotheses are contrasted: consistent θ, two θ’s (0.231 for intervals I–II, 0.481 for intervals III–VII), three θ’s (0.251 for intervals I–II, 0.714 for interval III, 0.400 for intervals IV–VI), four θ’s (0.231 for intervals I–II, 0.714 for interval III, 0.346 for interval IV, 0.426 for intervals V–VI). See Table 3 for additional information.
The test for consistent sampling intensity described above presents the logic for testing the distribution of sampling intensities across phylogeny. Instead of determining the number of sampling intensities within a single lineage, we now need to determine the number of sampling intensities within a clade. The likelihood of the simplest (one-parameter) hypothesis is simply a variant of equation (8):

\[ L[\hat{\theta} | n_1 \ldots n_X, R_1 \ldots R_X] = k \cdot \prod_{a=1}^{a=X} L[\hat{\theta} | n_a, R_a]. \]

Here \( n_a \) and \( R_a \) give the number of finds and possible finds for taxon \( a \), and \( \hat{\theta} \) is the estimated sampling intensity for all species in a clade. The most likely sampling intensity for a group of taxa (ignoring ghost taxa and lineages) is

\[ \hat{n} = \sum_{a=1}^{a=X} n_a - 2 \]
\[ \hat{R} = \sum_{a=1}^{a=X} R_a - 2 \]

for all \( X \) taxa with \( n \geq 2 \). \( (10) \)

Note that equation (10) includes taxa with \( n = 2 \). The sampling intensities of such taxa are 0 by equation (4), but we now need to assess the likelihood of a phylogenetically hypothesized sampling intensity given zero finds over some range.

Just as different rates of morphologic and molecular change can diagnose subclades and paraclines within a larger clade, different sampling intensities might diagnose subclades and paraclines. The likelihood of a sampling-intensity hypothesis (i.e., one positing different sampling intensities for \( Y \) different clades or paraclines) is simply a variant of equation (9):

\[ L[\hat{\theta}_1 \ldots \hat{\theta}_Y | n_1 \ldots n_X, R_1 \ldots R_X] \]
\[ = k \cdot \prod_{b=1}^{b=Y} \prod_{a=1}^{a=X_b} L[\hat{\theta}_b | n_{b,a}, R_{b,a}] \]

Now \( n_{b,a} \) and \( R_{b,a} \) give the number of finds and possible finds for taxon \( a \) within clade/paracline \( b \), and \( \hat{\theta}_b \) is the estimated sampling intensity for a subclade/paracline \( b \). Again, the single-sampling-intensity hypothesis is simply a special case of the multiple-sampling-intensity hypothesis where \( Y \) is one. Also, just as equation (9) is maximized when each interval has a unique sampling intensity, likelihood here will be maximized when each lineage has a unique sampling intensity.

Consider a clade with five taxa that have been sampled two or more times (Fig. 5). Given 40 finds over 255 possible finds, \( \theta_{\text{ALL}} = 0.125 \) (eq. 10). However, the probability of observing the known stratigraphic data for all five species under \( \theta_{\text{ALL}} \) is very low \( (p = 5.6 \times 10^{-4}; \) variation on eq. 8). Note that the calculations assume a Poisson distribution because phylogenetically implied gaps are treated as null hypotheses (see Wagner 1995). One derived subclade stands out with low \( n \)'s given observed \( R \)'s. This suggests a two-parameter hypothesis, with a primitive \( \hat{\theta}_1 = 0.195 \) for taxa 1, 4, and 5 and a derived \( \hat{\theta}_2 = 0.011 \) for taxa 2 and 3. The probability of the observed data \( (p = 1.9 \times 10^{-4}; \) variation on eq. 9) improves drastically under this hypothesis. The \( \Lambda \)-test decisively rejects the one-parameter hypothesis in favor of a two-parameter one \( (\Lambda = 20.8, p = 5.1 \times 10^{-5}) \). However, the data fail to reject the two-parameter model in favor of three-, four- or five-parameter hypotheses.

The likelihoods of a ghost taxon or lineage within the derived clade should be evaluated using equation (5) with \( n_a = 1 \) (i.e., \( [2 - 2] \) + \( [3 - 2] \)) and \( R_2 = 93 \) (i.e., \( [30 - 2] + [60 - 2] \)) test both hypotheses about durations and \( \hat{\theta}_2 \). One should evaluate ghost taxa and lineages within the rest of the larger clade using \( \hat{n} \) and \( \hat{R} \). The phylogeny is ambivalent concerning the appropriate \( \hat{\theta} \) for the ghost taxon linking the derived clade to the rest of the tree. However, it is most conservative to use \( \hat{n} \) and \( \hat{R} \) implying the minimum of two \( \hat{\theta} \)'s (i.e., \( \hat{\theta}_1 \)).

**Example 1: Lophospirid Gastropod Phylogeny**

Phylogenetic interpretations of cladograms are logical abductions (inferences) that assume synapomorphies (congruences) are homologies (Sobier 1988). Hypothesized topologies necessarily imply latest possible divergence times and thus include hypotheses about true durations. The implied durations permit the “if homologous then congruent”
FIGURE 5. Sampling intensities ($\theta$) for clades and unsampled ancestors ("ghost taxa") implied by estimated phylogeny and observed sampling. Observed sampling rejects a single-parameter hypothesis $\hat{\theta}_{\text{ALL}}$ in favor of a two-parameter hypothesis in which $\hat{\theta}_1$ applies to the paraclade including taxa 1, 4, and 5, and $\hat{\theta}_2$ applies to the clade including taxa 2 and 3. $n_i$ gives the number of finds for taxon $i$ whereas $R_i$ gives the number of possible finds within the stratigraphic range of taxon $i$. $n_{a.b}$ gives the finds for taxon $a$ in sampling group $b$; thus $n_{1.3} = 9$ (for taxon 5) whereas thus $n_{2.1} = 2$ (for taxon 2). However, the observed data fail to reject a two-parameter hypothesis in favor of more complicated ones.

$\hat{\theta}_{\text{ALL}} = 0.125 \quad \ln \prod_{a=1}^{b=5} L[\hat{\theta}_{\text{ALL}} | n_{a.}, R_a] = -19.0

\hat{\theta}_1 = 0.195 \quad \ln \prod_{b=2}^{a=X_b} \prod_{a=1}^{b=5} L[\hat{\theta}_1 | n_{b.a}, R_{b.a}] = -8.6

X_1 = 3 (Taxa 1, 4, 5) \quad X_2 = 2 (Taxa 2, 3)

$\Lambda = 20.8; \quad p = 5.1 \times 10^{-6}$

assumption (Patterson 1982) to be relaxed and cast as a testable hypothesis: if (if homologous then congruent), then stratigraphic data will be consistent with implied durations.

Wagner (1995, 1999) previously used stratigraphic data to test phylogenetic hypotheses for lophospirid gastropods, but without using the protocol suggested here. Those analyses evaluated likelihoods of phylogenetically implied range extensions from over 800 Ordovician-Silurian localities with lophospirid, murchisonioid, or etomarioid gastropods. Like lophospirids, most murchisonioids and etomarioids possessed aragonitic shells. The common co-occurrences of murchisonioids, etomarioids, and lophospirids suggest similar environmental preferences and tolerances. Eotomarioids and murchisonioids therefore offer a control for both taphonomy (Bottjer and Jablonski 1988) and paleoenvironment (see Holland 1995). Using taphonomic/environmental controls means that there are two results: samples that might include a species and do, and samples that might include a species (as indicated by the presence of similar fossils) but do not. Fossiliferous horizons lacking gastropods are not considered results of either sort and thus do not offer evidence against a hypothesized duration.

Using etomarioids and murchisonioids to tally the total number of plausible finds (rath-
er than using only the subset yielding lophospiroids as used by Wagner (1995)) increases the likelihoods of hypothesized gaps. Control taxa have little effect when assessing Ordovician lophospiroids because very few gastropod-bearing horizons lack lophospiroids. However, lophospiroids are much less common from Silurian localities. Adding murchisonioid and eotomarioid localities therefore increases gap likelihoods for Silurian taxa.

Hypothesized gaps implicit to two possible phylogenies are examined. One estimate is derived from minimum-steps parsimony (Edwards and Cavalli-Sforza 1964; Kluge and Farris 1969). The second is derived from a maximum-likelihood method (Wagner 1998, 1999) that assesses the likelihood of hypothesized amounts of change given observed character matrix structure and hypothesized amounts of stratigraphic gaps given observed $n$ and $R$. Relevant character data and analyses of character data are discussed elsewhere (Wagner 1999). Character and stratigraphic data can be accessed at http://pjw3.fmnh.org/MLDurations.html.

Tests reject consistent sampling intensities for seven species, five of which induce ghost lineages/taxa on the minimum-steps (hereafter MS) tree. In four such cases, sampling intensity for the first interval was lower than the global sampling intensity for a species. The 64 species known from 21 horizons determine the phylogenetically implied $\bar{G}$ of each range extension. The 64-parameter hypothesis has a log likelihood of -110, which means that an additional 45 parameters add very little predictive power. The maximum-likelihood (hereafter ML) tree invokes one more sampling intensity than does the MS tree, with a slightly greater likelihood of -109, which means that an additional 45 parameters add very little predictive power. The maximum-likelihood (hereafter ML) tree invokes one more sampling intensity than does the MS tree, with a slightly greater likelihood of $-110 - 109 = -107$.

Figure 6 shows the distribution of sampling intensities and implied range extensions across the MS tree. The figure caption gives each $\bar{G}$ as well as the $n$ and $R$ used to determine those $\bar{G}$'s. The caption also gives the hypothesized gap ($G$) of each range extension.

Tree likelihoods initially use model sampling intensities derived from empirical estimates and test range extensions for observed species (i.e., ghost lineages) (Table 4). Subsequent analyses test hypotheses about consistent sampling intensity over time before using sampling as a model. The final analyses assess likelihoods of both durations and sampling intensities. Range extensions are ghost taxa in all cases. This assessment is done by testing sampling intensities first among observed taxa and finally among all implied taxa (i.e., observed plus ghost taxa). Each modification increases the likelihood of the MS tree, with the final assessment increasing $L[MS \text{tree} | \text{stratigraphic data}]$ by several orders of magnitude relative to the initial estimates (Table 4). However, the data still reject the MS tree decisively in favor of the ML tree (the likelihood of which also increases markedly as analyses test increasing numbers of unknowns).

The topologic differences between the MS and ML trees (in part) increase the likelihood of the latter hypothesis given stratigraphic data. This renders the exact degrees of freedom problematic when evaluating the likelihood ratio test (Goldman 1993). However, even if one uses the number of differing branches (21) as degrees of freedom, then the data still soundly reject the MS tree in favor of the ML tree (Table 4). The ML tree also is more likely given observed character congruence than is the MS tree, even when likelihood tests and parsimony assumed the same model of character evolution (Wagner 1999). Thus, accommodating sampling inconsistencies does not change the conclusions drawn from older analyses.

Example 2: Seymour Island Ammonite Extinction

Numerous studies have tested hypotheses about the timing of the end-Cretaceous extinction event using quantitative methods. Several such studies have examined one particular data set, Macellari’s (1986) Late Campanian–Maastrichtian ammonites from Seymour Island (e.g., Strauss and Sadler 1989; Marshall 1995a; Solow 1996). Marshall (1995a) used ten ammonite species to test a hypothesized association between ammonite extinctions and the iridium-rich terminal Cretaceous beds. Marshall and Ward (1996) sug-
FIGURE 6. Favored hypothesis for distribution of sampling intensities (θ’s) across minimum-steps parsimony estimate of lophospiroid phylogeny. Light-gray branches denote those implying range extensions. Data reject the most likely 18-parameter hypothesis in favor of the 19-parameter hypothesis illustrated here; however, the same data fail to reject the 19-parameter hypothesis in favor of more complicated ones. Likelihoods of hypothesized unsampled ancestors are evaluated by the θ implied by the tree; in cases where one of two θ’s are possible, the lower θ is used. If more than two are possible, then the median θ is used. Nodal sampling parameters used to test ghost taxon hypotheses are: 1: θ = 0.143, n = 114, R = 800; 2: θ = 0.024, n = 6, R = 252; 3: θ = 0.021, n = 18, R = 880; 4: θ = 0.025, n = 94, R = 373; 5: θ = 0.078, n = 4, R = 53; 6: θ = 0.022, n = 94, R = 373; 7: θ = 0.018, n = 5, R = 285; 8: θ = 0.076, n = 6, R = 85; 9: θ = 0.045, n = 21, R = 469; 10: θ = 0.167, n = 6, R = 36; 11: θ = 0.044, n = 16, R = 361; 12: θ = 0.095, n = 4, R = 42; 13: θ = 0.032, n = 11, R = 341; 14: θ = 0.125, n = 7, R = 56. (n = Σ[ni − 2] and R = Σ[Ri − 2] for all clades/parACLades). Tested durations and implicit gaps (G) in numbers of sampling opportunities are (with applied sampling parameters in parentheses): A: G = 32 (1); B: G = 20 (2); C: G = 289 (2); D: G = 18 (3); E: G = 460 (3); F: G = 460 (3); G: G = 199 (3); H: G = 28 (3); I: G = 28 (3); J: G = 78 (3); K: G = 22 (1); L: G = 201 (8); M: G = 206 (8); N: G = 1 (6); O: G = 66 (6); P: G = 102 (7); Q: G = 102 (7); R: G = 160 (7); S: G = 203 (7); T: G = 81 (7); U: G = 60 (7); V: G = 309 (7); W: G = 78 (9); X: G = 55 (9); Y: G = 48 (9); Z: G = 100 (9); a: G = 103 (9); b: G = 50 (9); c: G = 36 (9); d: G = 50 (9); e: G = 36 (9); f: G = 41 (9); g: G = 4 (9); h: G = 22 (11); i: G = 60 (11); j: G = 156 (11); k: G = 330 (12); l: G = 251 (12); m: G = 63 (13); n: G = 38 (13); o: G = 251 (13).

suggested that there was high extinction among European ammonites before the terminal Cretaceous event. Here I test hypotheses about the number of extinction events among Maastrichtian ammonites.

The simplest explanation for a set of extinctions is that they reflect a single event. Conversely, the most likely hypothesis is that each species went extinct immediately after its last appearance. This invokes the maximum number of possible extinction events and thus the maximum number of parameters. The fossil record will tend to portray sudden mass extinctions as protracted events even when sampling is good (Signor and Lipps 1982), which might appear to bias likelihood tests in favor of suggesting protracted extinctions. However, such hypotheses must explain the data significantly better than hypotheses of few extinctions. Seymour Island ammonites’ last appearances occur at six “horizons” (“horizons” here reflect bins used in Macellari 1986: Fig. 5 and Marshall 1995a: Fig. 2). Although multiple finds per bin are possible, the presented data show only the presence/absence of each species in each bin. Also, the historical “null” hypothesis is that extinction events occur over extended periods of time. Because binomial distributions give lower likelihoods to gaps, the likelihoods assume a binomial distribution here.

Tests presented above reject consistent sam-
Table 4. Likelihoods of alternative estimates of lophospiroid phylogeny (from Wagner 1999) as θ and/or test hypotheses are modified. In tests 1–2, θ is a model. In tests 3–4, a model sampling intensity (θ) is chosen after testing hypotheses about consistency within and among lineages based on numbers of finds (n) and numbers of possible finds (D). In test 5, θ is part of a multiparameter hypothesis in which likelihood is maximized using hypothesized numbers of possible finds (D). “H” is the total hypothesis implicit to an inferred phylogeny (including divergence times); “d” denotes data. Log-likelihood test results (Λ) are included, as are the degrees of freedom necessary to invoke before stratigraphy fails to reject the parsimony tree (MST) in favor of the likelihood tree (MLT). Tests 1–3 use equation (2a) to test ghost lineages; tests 4–5 use equation (3a) to test ghost taxa (see Norell 1993). “ns.” denotes data. Log-likelihood test results (Λ) are included, as are the degrees of freedom necessary to invoke before stratigraphy fails to reject the parsimony tree (MST) in favor of the likelihood tree (MLT). Tests 1–3 use equation (2a) to test ghost lineages; tests 4–5 use equation (3a) to test ghost taxa (see Norell 1993). “ns.” denotes data.

<table>
<thead>
<tr>
<th>Test</th>
<th>θ</th>
<th>MST</th>
<th>Λ</th>
<th>MLT</th>
<th>Needed df</th>
</tr>
</thead>
<tbody>
<tr>
<td>eq. (2a), ( \theta = \frac{n}{R} )</td>
<td>Model</td>
<td>2 ( \times 10^{-239} )</td>
<td>-543.6</td>
<td>5 ( \times 10^{-36} )</td>
<td>924.3</td>
</tr>
<tr>
<td>eq. (2a), ( \theta = \frac{n - 2}{R - 2} )</td>
<td>Model</td>
<td>8 ( \times 10^{-157} )</td>
<td>-354.6</td>
<td>8 ( \times 10^{-27} )</td>
<td>60.1</td>
</tr>
<tr>
<td>eq. (2a), ns. θ</td>
<td>Hyp/Mod</td>
<td>1 ( \times 10^{-116} )</td>
<td>-331.1</td>
<td>1 ( \times 10^{-21} )</td>
<td>53.0</td>
</tr>
<tr>
<td>eq. (3a), ( \bar{\theta} = 10 )</td>
<td>Hyp/Mod</td>
<td>5 ( \times 10^{-85} )</td>
<td>-203.3</td>
<td>8 ( \times 10^{-14} )</td>
<td>30.2</td>
</tr>
<tr>
<td>eq. (5), ( \theta = \frac{\sum (n - 2)}{\sum (D - 2)} )</td>
<td>Hypo</td>
<td>3 ( \times 10^{-54} )</td>
<td>-123.4</td>
<td>1 ( \times 10^{-9} )</td>
<td>20.4</td>
</tr>
</tbody>
</table>

Modified likelihood estimates by treating unknowns as hypotheses increases our ability to reject the hypothesis of a single extinction (Table 5). Nevertheless, the data still reject a single extinction hypothesis in favor of a two-phase extinction (Table 6) fairly decisively, which corroborates Marshall and Ward’s (1996) suggestion of pre-terminal Cretaceous extinctions.

In this case, failing to accommodate sampling-intensity hypotheses favors the historical null hypothesis. However, likelihood tests reject the single-extinction hypothesis in favor of the two-extinction hypothesis even when sampling intensity is modeled simply and incorrectly (Table 6). Thus, accounting for unknown sampling parameters does not change conclusions regarding ammonite extinctions.

**Conclusions**

Factors such as variable and uncertain sampling intensity and unknown ancestral conditions all act to confound probability and likelihood tests of paleontological hypotheses. However, factors themselves are testable hypotheses. In two examples presented here, treating unknowns as hypotheses increases the likelihood of hypotheses that imply numerous gaps, but not so drastically as to change conclusions drawn from simple tests. There is no guarantee that the same will be true for other case studies. In particular, if there are few sampling opportunities, then these methods will find much greater likelihoods for numerous long gaps than will simple methods. Thus, although there is no evi-


dence that simultaneously testing sampling hypotheses will change the conclusions of previous analyses, this should not be assumed. More importantly, the protocols presented here negate at least some criticisms of quantitative methods and results because the supposed “violations” of probability tests now represent tested hypotheses. Finally, and most importantly, these protocols help show that when hypotheses about evolutionary history are being tested, the inconsistency of the fossil record is not insurmountable.

Acknowledgments

M. McGwire and S. Sosa provided the example data in Table 1. Some of the arguments presented herein originally were developed for the Nature Online Debate “Is the fossil record adequate?” For critical comments and discussion, I thank D. Erwin, T. Felix, M. (S.H.R.) Foote, J. H. Heywood, and J. B. C. Jackson. S. Lyons and J. Marcot provided especially helpful suggestions for clarifying a few particularly turgid paragraphs. Stratigraphic and phylogenetic data used in this study can be accessed at: http://pjw3.fmnh.org/MLDuration.html.

Literature Cited


Table 6. Results of likelihood-ratio tests of extinction hypotheses for ammonites. Lower left gives L values with G assessed by a χ² distribution with degrees of freedom equal to the difference in posited extinction events.

<table>
<thead>
<tr>
<th>Extinctions</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>—</td>
<td>8.9 × 10⁻⁴</td>
<td>9.6 × 10⁻⁴</td>
<td>1.1 × 10⁻³</td>
<td>1.8 × 10⁻⁴</td>
</tr>
<tr>
<td>2</td>
<td>11.04</td>
<td>—</td>
<td>0.09</td>
<td>0.08</td>
<td>0.11</td>
</tr>
<tr>
<td>3</td>
<td>13.89</td>
<td>2.85</td>
<td>—</td>
<td>0.15</td>
<td>0.07</td>
</tr>
<tr>
<td>4</td>
<td>15.98</td>
<td>4.95</td>
<td>2.09</td>
<td>—</td>
<td>0.27</td>
</tr>
<tr>
<td>5</td>
<td>17.18</td>
<td>6.14</td>
<td>3.29</td>
<td>1.19</td>
<td>—</td>
</tr>
</tbody>
</table>

The Likelihood of a Hypothesized Gap after First/Last Find and Hypothesized θ (Assuming a Poisson Distribution)

Given an observed range (R), a hypothesized duration (D) predicts a gap in sampling (D – R). If there can be multiple finds per stratigraphic bin, then the likelihood of D and a sampling intensity (θ) is

\[ L[D, \theta | n, R] = k \frac{(D - R + 1)\theta \cdot e^{-\theta} \cdot (R - 2)\theta^{(R - 2)} \cdot e^{-\theta}}{(D - R + 1)} \]

\[ S[D, \theta | n, R] = \ln(k) + \ln(D - R + 1) + \ln(\theta) - (D - R + 1)\theta 
+ (n - 2) \cdot \ln(R - 2) + (n - 2) \cdot \ln(\theta) - (R - 2)\theta 
- \ln[(D + R - 1) \cdot (n - 2)!] 
- \ln(k) + \ln(D - R + 1) + (n - 1) \cdot \ln(\theta) 
- (D - 1)\theta + (n - 2) \cdot \ln(R - 2) 
- \ln[(D + R - 1) \cdot (n - 2)!]. \]

This function describes a support surface. When testing a particular hypothesis (D – R), the function gives a support curve for alternative hypotheses of θ (sampling intensity). The maximum-likelihood hypothesis of sampling intensity given a hypothesized (D – R) is given by the value of sampling intensity at which the slope of the curve (i.e., the first derivative) equals 0 (Edwards 1992: p. 72). Portions of the function that do not refer to sampling intensity do not figure in the partial derivative, so the pertinent equation reduces to

\[ (n - 1) \cdot \ln(\theta) - (D - 1)\theta. \]

Therefore,

\[ S[D, \theta | n, R] \frac{\partial}{\partial \theta} = \frac{n - 1}{\theta} = (D - 1). \]

The derivative equals 0 when

\[ \frac{n - 1}{\theta} = (D - 1) \text{ or } \frac{n - 1}{\theta} = 0. \]

Appendix 2

The Likelihood of a Hypothesized Gap and Hypothesized θ (Assuming a Binomial Distribution)

Given an observed range (R), a hypothesized duration (D) predicts a gap in sampling (D – R). If there can be multiple finds per stratigraphic bin, then the likelihood of D and a sampling intensity (θ) is

\[ L[D, \theta | n, R] = k \cdot e^{-\theta} \cdot \theta^{n-2} \cdot (n - 2)! \cdot (R - 2)^{n-2} \cdot \theta^{(R - 2)} 
\]

\[ S[D, \theta | n, R] = \ln(k) - (R - 2)\theta 
+ (n - 2) \cdot \ln(R - 2) 
+ (n - 2) \cdot \ln(\theta) - (R - 2)\theta 
- \ln[(n - 2)!]. \]

The relevant portion when derivating by θ is

\[ (n - 2) \cdot \ln(\theta) - (D - 2) \theta. \]

Thus,

\[ S[D, \theta | n, R] \frac{\partial}{\partial \theta} = \frac{n - 2}{\theta} - (D - 2) \text{ or } \frac{n - 2}{\theta} = D - 2. \]

Appendix 3

The Likelihood of a Hypothesized gap after First/Last Find and Hypothesized θ (Assuming a Binomial Distribution)

Given an observed range (R), a hypothesized duration (D) predicts one find (the first or last) followed by (D – R) misses. If there can be only one find per stratigraphic bin (i.e., presence/absence), then the likelihood of D and a sampling intensity (θ) is
L[D, θ | n, R] = k · (1 - θ)^n · \left(\frac{R - 2}{n - 2}\right) \cdot θ \cdot (1 - θ)^{n - 1}

S[D, θ | n, R] = \ln(k) + (D - R) \cdot \ln(1 - θ) + \ln\left(\frac{R - 2}{n - 2}\right)
+ (n - 2) \cdot \ln(θ) + (R - n) \cdot \ln(1 - θ)
= \ln(k) + (n - 1) \cdot \ln(θ) + (D - n) \cdot \ln(1 - θ)
+ \ln\left(\frac{R - 2}{n - 2}\right)

The relevant portion when deriving by θ is

\( (n - 1) \ln(θ) + (D - n) \ln(1 - θ). \)

Thus,

\( S[D, θ | n, R] \frac{∂}{∂θ} = \frac{n - 1}{θ} - \frac{D - n}{1 - θ} \)

which equals 0 when

\( \frac{n - 1}{θ} = \frac{D - n}{1 - θ}. \)

cross multiplying,

\( (n - 1)(1 - θ) = Dθ - nθ \)
\( n - nθ - 1 + θ = Dθ - nθ \)
\( n - 1 + θ = Dθ \)
\( n - 1 = Dθ - θ \)
\( n - 1 = θ(D - 1) \) or \( θ = \frac{n - 1}{D - 1} \)

Appendix 4

The Likelihood of a Hypothesized Gap and Hypothesized θ
(Assuming a Binomial Distribution)

Given an observed range \( R \), a hypothesized duration \( D \) predicts a gap in sampling \( D - R \). If there can be only one find per stratigraphic bin (i.e., presence/absence), then the likelihood of \( D \) and a sampling intensity \( θ \) is

\[ L[D, θ | n, R] = k \cdot (1 - θ)^n \cdot \left(\frac{R - 2}{n - 2}\right) \cdot θ \cdot (1 - θ)^{n - 1} \]

\[ S[D, θ | n, R] = \ln(k) + (D - R) \cdot \ln(1 - θ) + \ln\left(\frac{R - 2}{n - 2}\right) \]
+ (n - 2) \cdot \ln(θ) + (R - n) \cdot \ln(1 - θ)
= \ln(k) + (n - 1) \cdot \ln(θ) + (D - n) \ln(1 - θ)
+ \ln\left(\frac{R - 2}{n - 2}\right)

The relevant portion when calculating the partial derivative of \( θ \) is

\( (n - 1) \ln(θ) + (D - n) \ln(1 - θ). \)

Thus,

\[ S[D, θ | n, R] \frac{∂}{∂θ} = \frac{n - 1}{θ} - \frac{D - n}{1 - θ} \]

which equals 0 when

\( \frac{n - 1}{θ} = \frac{D - n}{1 - θ}. \)

cross multiplying,

\( (n - 1)(1 - θ) = Dθ - nθ \)
\( n - nθ - 1 + θ = Dθ - nθ \)
\( n - 1 + θ = Dθ \)
\( n - 1 = Dθ - θ \)
\( n - 1 = θ(D - 1) \) or \( θ = \frac{n - 1}{D - 1} \)