X. Sampling Models, 1: Introduction to homogeneous sampling models

1 Basic framework for homogeneous sampling models

1.1 \( r \) is the per-capita rate of sampling per lineage-million-years.

1.2 “Sampling” means the joint incidence of preservation, exposure, collection, identification, etc.

1.3 Generally consider the relevance only of extinction rate \( q \) (tacitly assume \( p \approx q \) and truncation effects negligible).

This restriction can be fairly easily relaxed.

1.4 Assume all rates temporally and taxonomically homogeneous (for now).

1.5 It will often be useful to deal with “dimensionless rates” and “dimensionless time”.

Let \( q = 1 \), express \( r \) in terms of multiples of \( q \), and express time in terms of multiples of \( 1/q \) (expected mean duration).
2 Long-term average properties of a set of taxa


---

2.1 For a lineage with duration $T$, model sampling as a Poisson process with parameter $rT$.

Thus:

- $\Pr(\text{lineage never sampled}) = e^{-rT}$ (Poisson probability of zero successes.)
- $\Pr(\text{lineage sampled at least once}) = 1 - e^{-rT}$

---

**Processes**

Origination, extinction, and preservation occur at characteristic rates. A proportion of taxa is preserved; durations are reduced to observed ranges.

**Data**

Occurrences may be recorded in continuous time or in bins of length $dt$. Observations may be unrestricted or confined to window of length $w$. Data may include all occurrences or first and last occurrences only.

**Goals**

Given conditions under which observations are made, estimate characteristic rates of processes and proportion of taxa preserved.

---

**Figure 1**. Hypothetical examples of taxonomic durations, stratigraphic ranges, and nature of data. Various combinations of the three aspects of data require different approaches for estimating taxonomic rates, preservation rate, and proportion of taxa preserved (see text). For each of six lineages, short bars denote true times of origination and extinction and long bars denote fossil occurrences. Lineage 1 originates and is first preserved prior to window of observation, terminates and is last preserved during window, has five occurrences recorded in continuous time, and has three occurrences recorded in discrete time. Lineage 2 originates before window, terminates during window, and is never preserved. Lineage 3 originates before window, terminates after window, is preserved during window only, has six occurrences recorded in continuous time, and has four occurrences recorded in discrete time. Lineage 4 originates and terminates during window and has two occurrences recorded in both continuous and discrete time. Lineage 5 originates during window, terminates after window, and is recorded as a single hit during window (in both continuous and discrete time). Lineage 6 originates and terminates after window, has two occurrences recorded in continuous time, but has only one occurrence recorded in discrete time.
• Pr(lineage sampled exactly once) = \( rTe^{-rT} \) (Poisson probability of exactly one success.)

• Pr(observed range = \( t \) [\( t > 0 \)]) = \( r^2(T-t)e^{-r(T-t)} \). This is actually a density, not a probability, and is derived as follows:

\[
f(t|T) = \int_0^{T-t} [re^{-ry}][re^{-r(T-y-t)}] dy,
\]

where

- the first term within the integral is the density function for the first sampling event from the beginning of the duration
- the second term is the density function for the first event from the end of the duration (\([T - y - t]\) is the distance from the end working backward)

and

- the upper limit of integration is the longest that a true duration (\( T \)) can be if the range is \( t \).

• If sampling rate varies, substitute \( e^{-\int_0^T r_x dx} \) for \( e^{-rT} \).

2.2 These probabilities are then integrated over the entire distribution of durations (exponential or otherwise).

2.2.1 Example with untruncated exponential distribution

• The probability that a randomly chosen lineage will never be sampled is equal to:

\[
\int_0^\infty qe^{-qT} \cdot e^{-rT} dT = q/(r + q).
\]

• Pr(sampled at least once) = \( \int_0^\infty qe^{-qT} \cdot (1 - e^{-rT}) dT = r/(r + q) \). This is the expected proportion of lineages sampled, which has also been referred to as paleontological completeness.

• Pr(sampled exactly once) = \( \int_0^\infty qe^{-qT} \cdot rTe^{-rT} dT = qr/(r + q)^2 \). These are the singletons.

• Pr(observed range = \( t \) [\( t > 0 \)]) = \( \int_0^\infty qe^{-qT} \cdot r^2(T-t)e^{-r(T-t)} dT = qr^2e^{-qt}/(r + q)^2 \). (NB: This is actually a density.)

• By the number of times sampled, we mean the number of distinguishable stratigraphic horizons at which it is sampled, not the actual number of fossils or localities. Thus, a singleton may be sampled numerous times, but if these are all within the limits of resolution, at a single time horizon, this counts as being sampled once.
2.3 If we want to focus on distribution of stratigraphic ranges of lineages that are actually sampled, then we normalize all the probabilities by overall probability of sampling \((r/(r+q))\) in the exponential case. Thus:

- \(\text{Pr}(\text{singleton}) = q/(r+q)\)
- \(\text{Pr}(\text{range}=t \ [t > 0]) = qre^{-qt}/(r+q)\) (NB: This is actually a density.)
- The non-singleton part of this distribution is exponential with parameter \(q\). Thus, even though sampling is complete and therefore ranges are truncated, if we ignore singletons we expect shape of distribution to reflect true extinction rate accurately.

*Figure 4. Expected distribution of durations and stratigraphic ranges for continuous-time model with infinite window of observation. Curves are based on \(q = 0.1\) and \(r = 0.1\). A, Density of durations. B, Density of ranges. C, Reverse cumulative distribution of ranges. Closed square in B and C indicates discontinuity at ranges of zero (species present at one horizon). Duration density, range density, and reverse cumulative range distribution are all log-linear with negative slope equal to extinction rate, \(q\). The preservation rate can also be estimated from the distribution of ranges (Appendix 4).*

- Alroy’s data on mammal species seem to bear this out.
2.4 Comments on excess of singletons

- Implication of these models is that the frequency of singletons is overestimated in raw stratigraphic range data.

- This is not to say that any individual singleton taxon is an artifact; some taxa in reality did endure just one time interval.

- Frequency of singletons depends on both $q$ and $r$; thus this frequency is a good guide to comparing $r$ of two or more groups only if the groups have (roughly) the same $q$.


  - Assuming similar $q$ for calcitic and aragonitic taxa, this implies similar $r$ for the two groups.

- In much of $(q, r)$ parameter space, the probability that a taxon was in reality singleton, given that it is sampled in only one interval, is comparatively low.
Discrete time model

- q (instantaneous rate per lineage per interval)
- Pr(truly confined to single interval)
- R (sampling probability per interval)
- Pr(sampled as singleton | sampled at least once)
2.5 Density distribution can also be modified to accommodate finite window of observation (truncation of record at start and end).

We did this previously when looking at taxon age distributions, but we glossed over the details.

- Key is to establish density function for times of origination and density for times of extinction, given times of origination.
- Let \( w \) be length of window of observation.
- Assume diversity at \( t = 0 \) is equal to \( n_0 = 1 \) (this does not matter; everything scales to \( n_0 \)).
- Total progeny in window of time \( w \):

\[
M_w = \int_0^w p e^{(p-q)t} \, dt \\
= 1 + pw \quad \text{if} \quad p = q \\
= \frac{q - pe^{(p-q)w}}{q - p} \quad \text{if} \quad p \neq q
\]

- Origination density at time \( x \):

\[
g(x) = \frac{pe^{(p-q)x}}{M_w} \quad \text{if} \quad x > 0 \\
= \frac{1}{M_w} \quad \text{if} \quad x = 0
\]

- Extinction density at elapsed time \( z \) after origin (assuming origin at time \( x \)):

\[
f(z) = q e^{-qz},
\]

where \( z \) can take on any value between 0 and \((w - x)\).

- Thus, for taxa extinct within interval, duration density is given by:

\[
h(T) = g(0)q e^{-qT} + \int_0^{w-T} g(x)q e^{-qT} \, dx
\]

where the first term is for taxa extant at the start of the window and the second term is for taxa that originate during the window.
• Adjustment for taxa still alive at end:

\[
Pr(\text{extant at end}|\text{orig. at } x) = e^{-q(w-x)},
\]

Thus, the density function for duration within window \( h(T) \), where \( T = (w - x) \), must be incremented by \( g(x)e^{-q(w-x)} \).

• To obtain distribution of observed stratigraphic ranges, the resulting distribution of durations within the interval is then degraded just as the infinite exponential distribution was:

\[
\begin{align*}
\Pr(\text{lineage never sampled}) &= \int_0^w h(T)(e^{-rT}) \, dT \\
\Pr(\text{lineage sampled at least once}) &= \int_0^w h(T)(1 - e^{-rT}) \, dT \\
\Pr(\text{lineage sampled exactly once}) &= \int_0^w h(T)Te^{-rT} \, dT \\
\Pr(\text{observed range} = t \, [t > 0]) &= \int_0^w h(T)r^2(T - t)e^{-r(T-t)} \, dT
\end{align*}
\]

---

**Figure 6.** Expected density of durations and stratigraphic ranges for continuous-time model with finite window of observation. In all cases, \( p = 0.1, q = 0.1, \) and \( r = 0.1 \). Closed square in duration density indicates discontinuity at within-interval durations equal to the interval length itself. Only species extant at both the start and end of interval have within-interval durations equal to the interval length. Closed square in range density indicates discontinuity at ranges of zero. A–C, Window of observation (interval) equal to 10, 20, and 40. Note that, if \( p = q \), as the window of observation becomes longer, the duration and range distributions approach those for an infinite window of observation (Fig. 4, but see Appendix 1).
3 Basic inverse problem: estimates of overall sampling rate per unit time

CAVEAT: Although often overlooked, it is important to bear in mind that most approaches estimate the rate of sampling for taxa of which we even have a record. If entire clades, environments, or geographic regions are unsampled, or nearly so, we can still end up with high numerical estimates of sampling rate even though overall sampling on average may be much poorer than suggested by these estimates.

3.1 Gap analysis (discrete time)

Overall $R = 5/16$.

$R_i(i = 1, ..., 6) : NA, 0, 1/4, 2/4, 2/4, NA$

Needless to say, in practice, you would not be likely to carry out estimates with so few observations.

<table>
<thead>
<tr>
<th>Species</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Opportunities for sampling</strong> (excluding range endpoints)</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>--</td>
<td>2</td>
<td>--</td>
<td>1</td>
<td>--</td>
<td>3</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td><strong>Occurrences</strong> (excluding range endpoints)</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>--</td>
<td>1</td>
<td>--</td>
<td>0</td>
<td>--</td>
<td>0</td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

3.1.1 Sampling rate (probability) for a set of taxa

- $R$ is the probability of sampling a lineage at least once in a time interval, assuming it ranges all the way through. Thus $R = 1 - e^{-r\Delta t}$, where $\Delta t$ is the interval duration.

- Let $t_i$ be the observed range of taxon $i$, and let $n_i$ be the number of distinct intervals in which it occurs.

- Estimate $R$ as $\sum_{t_i>2}(n_i - 2) / \sum_{t_i>2}(t_i - 2)$, where only taxa with $t > 2$ are considered.
• Intervals of first and last appearances are disregarded because they must be intervals in which the taxon is sampled. Including them leads to an overestimate of $R$.

3.1.2 Sampling rate (probability) for a single interval of time

• Estimate $R$ as $N_{bt, sampled}/N_{bt}$, where $N_{bt, sampled}$ is the number of lineages that range all the way through the interval and are sampled at least once within it.

• As with the estimate averaged over all time, taxa making a first and/or last appearance within the interval are disregarded.

3.2 Estimate long-term average rates from distribution of stratigraphic ranges in continuous time (with no information on internal occurrences between range endpoints)

• Let $f(t)$ be the density function (relative probability of stratigraphic range of $t$).

• $f(0) = q/(r + q)$

• $f(t) = qre^{-qt}/(r + q)$, $t > 0$

• Let $n_s$ be the number of singletons (those with range=0).

• Let $m$ be the number of non-singletons, with observed ranges $t_1, ..., t_m$.

• Then the likelihood function is:

$$L(q, r) = \prod_{i=1}^{m} f(t_i) \times f(0)^{n_s}$$

$$= \prod_{i=1}^{m} \frac{qre^{-qt_i}}{(r + q)} \times \left( \frac{q}{r + q} \right)^{n_s}$$

• and the corresponding support (log-likelihood) function is:

$$S(q, r) = \sum_{i=1}^{m} \ln(q) + \ln(r) - qt_i - \ln(r + q) + n_s[\ln(q) - \ln(r + q)]$$

$$= (m + n_s)\ln(q) + m\ln(r) - (m + n_s)\ln(r + q) - q \sum_{i} t_i$$
Since we are trying to solve for two parameters, we need to take the partial derivative with respect to each:

$$\frac{\partial S}{\partial q} = \frac{(m + n_s)}{q} - \frac{(m + n_s)}{(r + q)} - \sum_i t_i$$

$$\frac{\partial S}{\partial q} = 0 \Rightarrow \frac{r}{q(r + q)} = \frac{\sum_i t_i}{m + n_s} \quad (A)$$

and

$$\frac{\partial S}{\partial r} = \frac{m}{r} - \frac{(m + n_s)}{(r + q)}$$

$$\frac{\partial S}{\partial r} = 0 \Rightarrow r = \frac{qm}{n_s} \quad (B)$$

If we substitute the expression for \(r\) in (B) into (A) we ultimately end up with:

$$\hat{q} = \frac{m}{\sum_i t_i}$$

and

$$\hat{r} = \frac{\hat{q}(1 - f_0)}{f_0},$$

where the \(\hat{\cdot}\) denotes the maximum likelihood estimate and \(f_0\) is the proportion of observed taxa that are singletons, i.e \(f_0 = \frac{n_s}{n_s + m}\).

- Note that \(\hat{q}\) is simply the inverse of the mean observed range of non-singleton taxa. Example: planktic vs. nonplanktic gastropod durations (Hansen 1980, *Paleobiology* 6:193).

- Note also that \(r\) and \(q\) must be jointly estimated; the estimate of \(r\) depends on a prior estimate of \(q\). This is a rather common situation in parameter estimation.

We already saw that the expected probability of being sampled at least once is equal to \(r/(r + q)\). Therefore, once we have the estimates \(\hat{r}\) and \(\hat{q}\), we would estimate the proportion of taxa sampled at least once as \(\hat{r}/(\hat{r} + \hat{q})\).

NB: The particular expressions for \(f_t\) used above assume infinite window of observations. We can also use the more general expressions with finite window and solve (numerically) for \(p, q,\) and \(r\).
Raw mean range: 4.4 m.y. and 2.1 m.y. \((q = 0.23\) and 0.48 per Lmy))
Mean range without singletons: 5.3 m.y and 3.5 m.y. \((q = 0.19\) and 0.28 per Lmy)
Thus, planktic-nonplanktic difference holds, but reduced in magnitude.
Caveat: Estimate of \(1/q\) as mean range without singletons strictly holds only for continuous-time model, but is acceptable for binned data if bins are relatively short.

### 3.3 Estimate long-term average rates from distribution of stratigraphic ranges in continuous time (including information on internal occurrences between range endpoints)

Intuitively, it seems reasonable that we should obtain estimates of \(q\) and \(r\) with less uncertainty if we can incorporate additional data on the occurrences within ranges. This is what is done by Solow and Smith (1997, *Paleobiology* 23:271-277).

- Let there be \(N\) taxa, each known from \(n_i\) distinct stratigraphic horizons (for example \(n_i = 1\) would denote a singleton), and let \(d_i\) be the observed stratigraphic range for each.
- Then:

\[
\hat{r} = \frac{\left[\sum (n_i - 1)\right]^2}{\left[\sum d_i\right] \left[\sum n_i\right]} 
\]
and

\[ \hat{q} = N\hat{r}/\sum_i (n_i - 1) \]

(See Solow and Smith for the derivation of this.)

### 3.4 Estimate long-term average rates from distribution of stratigraphic ranges in discrete time (with no information on internal occurrences between range endpoints) (see Foote and Raup, 1996, *Paleobiology* 22:121-140).

#### 3.4.1 Basic forward model for discrete time

![Figure 1](image)

- **True duration**
  - 15 m.y.
  - 15 m.y.
  - 10 m.y.

- **Fossil range**
  - 7 m.y.
  - 7 m.y.
  - 1 m.y.

\[ R = \text{Prob(preservation at least once in interval)} \]
\[ (1 - R) = \text{Prob(non-preservation)} \]

In general, \( \text{Prob(range = t if duration = T)} \):

\[
\begin{align*}
&= (1 - R)^T \quad &\text{if } t = 0 \\
&= T(1 - R)^{T - 1}R \quad &\text{if } t = 1 \\
&= (T - t + 1)(1 - R)^{T - t}R^t \quad &\text{if } t > 1
\end{align*}
\]

**Figure 1.** Relationship between taxonomic duration and fossil range. Time is in discrete intervals (here, one million years), and durations and ranges are treated as ranging through each interval of occurrence. Empty intervals indicate non-preservation; hatched intervals indicate preservation of taxon at least once during interval. Probability that taxon with true duration 7 has range 1 is easily derived (Appendix 1).

- Let \( q \) be the per-capita rate of extinction per discrete time interval and let \( R \) be the sampling probability per interval, as defined above. Because we are using a discrete-time model, we tacitly assume that a taxon present in an interval is present throughout the interval.

- Let \( T \) denote the true duration and \( t \) the observed stratigraphic range. Note that, in contrast to the continuous-time case, a singleton is a taxon with \( t = 1 \) (regardless of
how many times it is sampled within the interval), and a taxon with \( t = 0 \) is one that is not sampled at all. Then, for a single taxon of duration \( T \):

\[
Pr(t = 0) = (1 - R)^T \\
Pr(t \geq 1) = 1 - (1 - R)^T \\
Pr(t = 1) = RT(1 - R)^{T-1}
\]

and

\[
Pr(t \mid t > 1) = (T - t + 1)R^2(1 - R)^{T-t}
\]

- Let \( P_D(T) \) be the probability that a duration is equal to \( T \) (strictly speaking, that it falls between \((T-1)\) and \(T\)), where time is tabulated in number of intervals. Thus, \( P_D(T) = e^{-q(T-1)} - e^{-qT} \).

- We now sum over the probability entire distribution of durations, each time considering the probability distribution of possible observed ranges (see Appendix 1 of Foote and Raup).

- Finally, we normalize by the probability of being sampled at least once (Foote and Raup Eq. 2).
Appendix 1

Predicting distribution of observed stratigraphic ranges from original distribution of taxonomic durations

Let $T$ be the original duration of some taxon, let $R$ be the probability that the taxon is preserved at least once in an arbitrary increment of time, and let $t$ be the preserved stratigraphic range (using discrete-time convention; see Fig. 1). If $R$ is stochastically constant through time, the probability that a taxon with original duration $T$ is not preserved at all (i.e., the probability that $t = 0$) is equal to $(1 - R)^T$, and therefore the probability that it is preserved at least once (i.e., the probability that $t \geq 1$) is equal to

$$P(t) = 1 - (1 - R)^T. \quad (1)$$

Let there be an arbitrary distribution of taxonomic durations, where $P_n(T)$ is the probability that the duration equals $T$ (strictly speaking, the probability that it is between $T - 1$ and $T$, since we are dealing with discrete time intervals). For example, with constant extinction rate, $P(T)$ is equal to $e^{- \sigma T}$ (Appendix 2A). If the probability of preservation is homogeneous among taxa, then the probability that a taxon randomly chosen from this distribution is preserved at least once is equal to

$$P = \sum_{T=1}^{\infty} P(T)P_n(T). \quad (2)$$

This is, of course, the proportion of taxa preserved (completeness).

We desire a general expression to relate $T$, $t$, and $R$. If the original duration is $T$ and the range, $t$, is equal to 1, this single occurrence can be in any of $T$ places stratigraphically, and there must be $T - 1$ intervals of non-preservation. Thus, the probability that the range equals 1, given that the duration equals $T$, is given by

$$P(t=1, T) = TR(1 - R)^{T-1} \quad (3)$$

For ranges greater than unity, there must be $T - t$ instances of non-preservation outside the fossil range, and there must be two instances of preservation at the endpoints; these endpoints can fall in $(T-x + 1)$ different ways (see Fig. 1). The number of instances of preservation between the endpoints can be anything from zero to $T-2$, but all the corresponding probabilities sum to unity, and so can be ignored. Thus, the probability that the range is equal to $t$, given that the duration is equal to $T$, is given by

$$P(t, T) = (T - t + 1)R^t(1 - R)^{T-1}. \quad (4)$$

Equations (3) and (4) give the probability that a certain range will be observed, given a particular original duration. We would like to know the probability of a certain range summed over the entire distribution of original durations. Moreover, since we only observe taxa that are preserved, we would like to normalize this expression by the overall probability of being preserved at least once (eq. 2). We therefore have the relative frequency of fossil ranges (the probability, for preserved taxa, that the fossil range is equal to $t$):

$$f(t) = \frac{\sum_{T=1}^{\infty} P(t, T)P_n(T)}{P}. \quad (5)$$

From this expression, we can calculate the mean and median fossil range for some initial distribution of original durations (e.g., Fig. 2B). From equations (1) and (2), it follows that the probability that a preserved taxon has original duration $T$, given that it is preserved at least once, is equal to

$$g(T) = P(T)P_n(T)/P. \quad (6)$$

This expression allows us to calculate the mean and median original duration of taxa that are preserved at least once (e.g., Fig 2B).
3.4.2 Principal results of the forward model

1. The distribution of ranges, ignoring singletons, is exponential with parameter \( q \). Thus we can estimate \( q \) through regression of \( \ln[f_t] \) against \( t \) (where \( f_t \) is the proportion of observed taxa with range equal to \( t \)), for all \( t > 1 \).

2. \( R = f_2^2/(f_1 f_3) \). This ratio, the so-called \( \text{FreqRat} \), provides an estimate of the sampling probability per taxon per time bin.

![Diagram](image)

**Figure 2.** Analytic degradation of initially exponential distribution of taxonomic durations, corresponding to extinction rate \( q \) of 0.1 per taxon per million years, and preservation probability \( R \) of 0.2 per taxon per million years. Frequencies in this and all subsequent figures are raw values, not cumulative frequencies. A. Original durations of all taxa. B. Original durations of the subset of taxa preserved at least once (in this case, 72% of taxa). Because shorter-lived taxa are less likely to be preserved, fossil taxa represent a biased subset having longer durations on average. C. Fossil ranges of preserved taxa. Because of truncation, the median range is shorter than the true median duration. If single hits are ignored, the remaining range-frequency relationship is log-linear with slope equal to \( -q \) (Appendix 2). Thus, fossil ranges can be used to estimate the true extinction rate. The range-frequency ratio \( \text{FreqRat} \), i.e., \( R(2)/R(1)R(3) \), is exactly equal to the probability of preservation, \( R \) (Appendix 3). Thus, fossil ranges can be used to estimate preservation potential and completeness.
3.4.3 Values of $f_t$ can of course be adjusted for finite window of observation, as outlined above for continuous-time model.

3.4.4 Problem with the foregoing estimates of $q$ and $R$: It turns out they are biased at finite sample sizes. This can be shown either by simulation or analytically, as follows:

- Let $f_1$, $f_2$, and $f_3$ be the true probabilities that a taxon will have a range of 1, 2, or 3 intervals, and let $f_4 = 1 - f_1 - f_2 - f_3$ be the probability that the range is greater than 3 intervals. These probabilities come from section (4) above.

- Let $k_1$, $k_2$, and $k_3$ be the number of taxa with ranges of 1, 2, and 3 intervals. Let $k_4$ be the number of all the rest. Let $N$ be the total number of taxa ($N = k_1 + k_2 + k_3 + k_4$). For a given dataset, the $FreqRat$ would be calculated as $k_2^2/(k_1k_3)$.

- The numbers $k_1,...,k_4$ follow a multinomial distribution:

$$Pr(k_1, k_2, k_3, k_4) = \frac{N!}{k_1!k_2!k_3!k_4!} f_1^{k_1} f_2^{k_2} f_3^{k_3} f_4^{k_4}$$

Likewise,

$$f_4 = (e^{-R}/P_4)A$$

and so on. Therefore, for $t > 1$,

$$\ln(f(t)) - \ln(f(t + 1)) = q.$$
The expected value of the FreqRat is given by:

\[
E\left(\frac{k_2}{k_1k_3}\right) = \frac{\sum_{k_1=1}^{N} \sum_{k_2=1}^{N-k_1} \sum_{k_3=1}^{N-k_2-k_1} k_2^2 k_1 k_3 \cdot Pr(k_1, k_2, k_3, k_4)}{\sum_{k_1=1}^{N} \sum_{k_2=1}^{N-k_1} \sum_{k_3=1}^{N-k_2-k_1} Pr(k_1, k_2, k_3, k_4)},
\]

where the sum is taken only over those cases where \(k_1\), \(k_2\), and \(k_3\) are positive, and the denominator normalizes by the sum of probabilities over these cases.

Carrying out these tedious calculations, we find that the expected FreqRat is an overestimate of the true value of \(R\) at finite sample sizes, and the bias is greater as \(N\) is smaller.

3.4.5 A solution to the problem is to use maximum likelihood to jointly estimate (numerically) the parameters \(q\) and \(R\).

Let \(f_t(q, R)\) be the probabilities predicted for a given pair of \((q, R)\) and let \(k_t\) be the observed frequencies. Then the likelihood is given as

\[
L(q, R|k_t) = \prod_{t=1}^{\infty} f_t(q, R)^{k_t}
\]

and the support is

\[
S(q, R|k_t) = \sum_{t=1}^{\infty} k_t \ln[f_t(q, R)]
\]

Please note that \(k_t\) in the foregoing is the same as \(d_x\) from our dynamic survivorship analysis!

There seems to be no analytical solution, so we explore values of \(q\) and \(R\) until we find the pair that maximizes the likelihood (or support). These are then our maximum-likelihood estimators \(\hat{q}\) and \(\hat{R}\).

3.5 Taxonomic heterogeneity in sampling probability

3.5.1 Overall estimate of \(r\) or \(R\) weighted more heavily by better-sampled component of fauna.

3.5.2 Some model calculations suggest estimate of \(q\) may not be seriously distorted.
number of initial species (not necessarily the same number of preserved species). One class has a constant preservation probability of 0.5, while that of the other varies. For each case, 100 species are randomly sampled from the expected distribution of ranges, and parameters are estimated from ranges by maximum likelihood using the assumption of homogeneous preservation. This simulation is repeated 500 times to yield the mean estimate (closed square) and its standard deviation (error bar). Dashed lines show true parameter values. A, Preservation probability. Dashed line shows simple mean of the two preservation probabilities. Open squares show the mean of the two preservation probabilities weighted by the completeness of the respective class. B, Completeness. C, Extinction rate. Preservation probability and completeness are overestimated if taxa differ markedly in preservability, but extinction rate is rather accurately estimated.
3.6 Comparing estimates from various approaches

3.6.1 Relative strengths of these and other approaches have yet to be thoroughly evaluated.

3.6.2 Preliminary comparisons suggest that they are not greatly out of line with each other.

3.6.3 Methods that use all occurrences (not just range endpoints) will generally yield more precise estimates of $r$ and $R$. 
### Table 1. Estimation of extinction rate, preservation rate (or probability), and completeness for 101 Cambro-Ordovician trilobite species (Stitt 1977). Numbers in parentheses are limits of three-unit support region.

<table>
<thead>
<tr>
<th>Method</th>
<th>Extinction rate</th>
<th>Preservation rate (or probability)</th>
<th>Completeness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous time, infinite window, ranges alone used, slope of reverse cumulative frequency</td>
<td>0.0077 per foot</td>
<td>0.021 per foot</td>
<td>0.73</td>
</tr>
<tr>
<td>Continuous time, infinite window, all occurrences used, maximum likelihood (Solow-Smith method)</td>
<td>0.0086 (0.0067, 0.0109) per foot</td>
<td>0.047 (0.042, 0.053) per foot</td>
<td>0.85 (0.81, 0.88)</td>
</tr>
<tr>
<td>Continuous time, infinite window, ranges alone used, maximum likelihood</td>
<td>0.076 (0.0057, 0.0099) per foot</td>
<td>0.022 (0.012, 0.042) per foot</td>
<td>0.74 (0.63, 0.84)</td>
</tr>
<tr>
<td>Continuous time, infinite window, ranges alone used, maximum likelihood</td>
<td>0.0069 (0.0051, 0.0092) per foot</td>
<td>0.022 (0.012, 0.042) per foot</td>
<td>0.74 (0.65, 0.85)</td>
</tr>
<tr>
<td>Discrete time, infinite window, ranges alone used, method of Foote and Raup (1996)</td>
<td>0.35 per 60 feet</td>
<td>0.73 per 60 feet</td>
<td>0.9</td>
</tr>
<tr>
<td>Discrete time, infinite window, ranges alone used, maximum likelihood</td>
<td>0.42 (0.30, 0.55) per 60 feet</td>
<td>0.65 (0.35, 0.95) per 60 feet</td>
<td>0.85 (0.64, 0.98)</td>
</tr>
<tr>
<td>Discrete time, finite window, ranges alone used, maximum likelihood</td>
<td>0.39 (0.30, 0.50) per 60 feet</td>
<td>0.64 (0.35, 0.95) per 60 feet</td>
<td>0.84 (0.64, 0.98)</td>
</tr>
</tbody>
</table>

### Table 2. Estimation of extinction rate (per 0.7 m.y.), preservation probability (per 0.7 m.y.), and completeness for 2941 Cenozoic mammal species (Alroy 1994). Numbers in parentheses are limits of three-unit support region. Cretaceous and extant species are omitted.

<table>
<thead>
<tr>
<th>Method</th>
<th>Extinction rate</th>
<th>Preservation probability</th>
<th>Completeness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discrete time, infinite window, ranges alone used, method of Foote and Raup (1996)</td>
<td>0.28</td>
<td>0.25</td>
<td>0.63</td>
</tr>
<tr>
<td>Discrete time, infinite window, ranges alone used, maximum likelihood</td>
<td>0.33 (0.32, 0.35)</td>
<td>0.37 (0.34, 0.40)</td>
<td>0.67 (0.65, 0.70)</td>
</tr>
<tr>
<td>Discrete time, finite window, ranges alone used, maximum likelihood</td>
<td>0.29 (0.30, 0.34)</td>
<td>0.37 (0.34, 0.40)</td>
<td>0.67 (0.65, 0.70)</td>
</tr>
</tbody>
</table>
FIGURE 9. Maximum likelihood estimation of \((q,r)\) for Cambro-Ordovician trilobites, continuous-time model.

A. All occurrences used, infinite window of observation assumed (Solow and Smith 1997). B. Ranges alone used, infinite window of observation assumed. C. Ranges alone used, finite window assumed. In each panel, the point marks the maximum likelihood estimator \((\hat{q}, \hat{r})\). The inner white zone contains all values of \((q,r)\) within three units of support of \((\hat{q}, \hat{r})\) (i.e., approximate 95% confidence region). The next black band contains values of \((q,r)\) between three and five units of support, and the
Figure 8. Comparison of estimation procedures with all occurrence data and with range endpoints only, for simulations scaled to \((q,r) = (0.0086, 0.047)\) (based on trilobite data; Table 1). Simulations use continuous-time model with an infinite window of observation to generate ranges and number of occurrences. Each point represents maximum likelihood estimation of \((q,r,P_r)\) for a single simulation of 100 species (100 simulations total). Error bars show mean ± one standard deviation. The
3.7 Correlations among parameter estimates

- Recall that $q$ and $r$ are jointly estimated.
- Increasing $q$ reduces stratigraphic ranges, as does decreasing $r$.
- Thus it makes sense that an over- or underestimate of $q$ could be “compensated for” by a corresponding over- or underestimate of $r$.
- This seems to be borne out by shape of contours in likelihood surface.
- And it can also be seen in estimates from simulated data.

![Figure 12](image-url)  
*Figure 12. Correlation between estimates of $q$ and $r$ in simulations of Figure 11. Correlation coefficient is equal to 0.65 using ranges only versus 0.55 using all occurrences. Compare with Figures 10 and 11.*
4 Estimate of proportion of taxa sampled at least once.

4.1 Continuous-time model (assuming infinite window)

\[ \hat{P}_p = \hat{r} \frac{\hat{r}}{\hat{q} + \hat{r}} \]

This can be modified for finite window of observation.

4.2 Discrete-time model (assuming infinite window)

\[ \hat{P}_p = \sum_{T=1}^{\infty} \left[ e^{-\hat{q}(T-1)} - e^{-\hat{q}T} \right] \left[ 1 - (1 - \hat{R})^T \right], \]

where \( \hat{q} \) is the extinction rate scaled per interval rather than per m.y. This can also be modified for finite window of observation.

4.3 Comparison between number of known fossil taxa and estimated total progeny

(e.g. Foote, 1996, Paleobiology 22:141-151, Table 1.)

---

**Table 1.** Estimated total progeny (\( M \)), proportion of species preserved (\( P_p \)), and probability of finding direct ancestor-descendant pairs (\( P_a \)) for three different models of global Phanerozoic diversity of the paleontologically important marine groups, assuming budding birth-death model and homogeneous preservation with probability corresponding to the \( \hat{r} \) value of \( P_p \).

<table>
<thead>
<tr>
<th>Extinction rate (per species per m.y.)</th>
<th>Median species duration (m.y.)</th>
<th>Deterministic, constant diversity</th>
<th>Deterministic, linear diversity increase</th>
<th>Stochastic, exponential diversity increase*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>( M )  ( P_p )  ( P_a )</td>
<td>( M )  ( P_p )  ( P_a )</td>
<td>( M )  ( P_p )  ( P_a )</td>
</tr>
<tr>
<td>0.05  13.8  4.1 \times 10^5  0.049  0.081</td>
<td>2.0 \times 10^4  0.098  0.15</td>
<td>5.2 \times 10^6  0.38  0.41</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.10  6.9  8.1 \times 10^5  0.025  0.043</td>
<td>4.1 \times 10^4  0.049  0.081</td>
<td>9.3 \times 10^6  0.22  0.37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.25  2.8  2.0 \times 10^7  0.010  0.019</td>
<td>1.0 \times 10^7  0.020  0.036</td>
<td>2.4 \times 10^8  0.083  0.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.50  1.4  4.1 \times 10^9  0.0049  0.0091</td>
<td>2.0 \times 10^7  0.0059  0.019</td>
<td>4.9 \times 10^9  0.041  0.070</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.0   0.7  8.1 \times 10^9  0.0025  0.0048</td>
<td>4.1 \times 10^9  0.0049  0.0091</td>
<td>1.1 \times 10^10  0.018  0.033</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Strictly speaking, this diversity model is not exponential. For a given origination and extinction rate, the expected diversity over time follows an exponential trajectory if this expectation includes the possibility of prior extinction of the clade in question (see Kendall 1948; Raup 1985). The expected diversity of life over the Phanerozoic, assuming survival of life is in the Recent, resembles exponential growth somewhat but increases extremely rapidly early on. This is intuitively reasonable, as clades that persist for a long time are likely to be those that get off to a good start.

1 The expected total progeny of this stochastic process is based on Raup (1985, eq. A20). This equation is problematic, as the rationale behind it is not immediately obvious (D. M. Raup personal communication 1994). I have found, by comparing results of the equation to simulation results, that the equation is accurate when origination and extinction rates are roughly equivalent. However, with large differences between origination and extinction rates, the equation predicts values of total progeny that disagree with simulation results. The reader should therefore be cautioned against using this equation when rates of origination and extinction differ substantially from each other. With high-speed computers, it is a simple task to use simulation to determine whether the equation is greatly biased for given values of origination rate, extinction rate, and clade age.
4.4 Proportion of living taxa with a fossil record


![Figure 1](image-url) Absolute measures of completeness for higher taxa of marine animals. For each taxon, the proportion of living families with a fossil record is compared with the estimated probability of genus preservation per stratigraphic interval. The data show a positive trend, indicating a correlation between the different completeness measures at two taxonomic levels and temporal scopes of data (genera mainly pre-Cenozoic; families mainly Cenozoic). Taxa that deviate from the trend are enclosed in boxes and analysed in Fig. 2. Two measures of genus-level completeness, reflecting different subdivisions of the Phanerozoic, are shown connected by a bar. The open squares denote the 103-interval timescale, whereas the closed squares denote the 107-interval timescale. The dotted diagonal line has unit slope.