

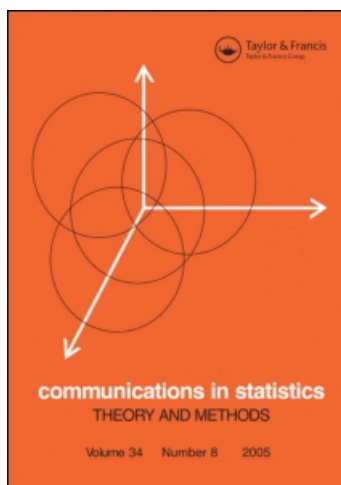
This article was downloaded by: [University of Texas at El Paso]

On: 20 April 2011

Access details: Access Details: [subscription number 933589298]

Publisher Taylor & Francis

Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



Communications in Statistics - Theory and Methods

Publication details, including instructions for authors and subscription information:

<http://www.informaworld.com/smpp/title~content=t713597238>

The Distribution of Family Sizes Under a Time-Homogeneous Birth and Death Process

Panagis Moschopoulos^a; Max Shpak^b

^a Department of Mathematical Sciences, University of Texas at El Paso, El Paso, Texas, USA ^b

Department of Biological Sciences, University of Texas at El Paso, El Paso, Texas, USA

Online publication date: 11 May 2010

To cite this Article Moschopoulos, Panagis and Shpak, Max(2010) 'The Distribution of Family Sizes Under a Time-Homogeneous Birth and Death Process', Communications in Statistics - Theory and Methods, 39: 10, 1761 – 1775

To link to this Article: DOI: 10.1080/03610920902898498

URL: <http://dx.doi.org/10.1080/03610920902898498>

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: <http://www.informaworld.com/terms-and-conditions-of-access.pdf>

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

The Distribution of Family Sizes Under a Time-Homogeneous Birth and Death Process

PANAGIS MOSCHOPOULOS¹ AND MAX SHPAK²

¹Department of Mathematical Sciences, University of Texas
at El Paso, El Paso, Texas, USA

²Department of Biological Sciences, University of Texas
at El Paso, El Paso, Texas, USA

The number of extant individuals within a lineage, as exemplified by counts of species numbers across genera in a higher taxonomic category, is known to be a highly skewed distribution. Because the sublineages (such as genera in a clade) themselves follow a random birth process, deriving the distribution of lineage sizes involves averaging the solutions to a birth and death process over the distribution of time intervals separating the origin of the lineages. In this article, we show that the resulting distributions can be represented by hypergeometric functions of the second kind. We also provide approximations of these distributions up to the second order, and compare these results to the asymptotic distributions and numerical approximations used in previous studies. For two limiting cases, one with a relatively high rate of lineage origin, one with a low rate, the cumulative probability densities and percentiles are compared to show that the approximations are robust over a wide range of parameters. It is proposed that the probability distributions of lineage size may have a number of relevant applications to biological problems such as the coalescence of genetic lineages and in predicting the number of species in living and extinct higher taxa, as these systems are special instances of the underlying process analyzed in this article.

Keywords Birth-death process; Genealogy; Hypergeometric function; Phylogeny; Taxon size.

Mathematics Subject Classification Primary 60J80; Secondary 62E15, 62E17, 62P10.

1. Introduction

The number of individuals in a genetic lineage (family size) and the formally equivalent problem of the number of species within genera or other higher taxa has been of wide interest to demographers and biologists alike (e.g., Anderson, 1974; Burlando, 1990; Watson and Galton, 1875; Yule, 1924). Empirically, the distributions of family sizes

Received May 30, 2008; Accepted March 16, 2009

Address correspondence to Max Shpak, Department of Biological Sciences, University of Texas at El Paso, El Paso, TX 79968, USA; E-mail: mshpak@utep.edu

in humans and taxon sizes in a number of organisms have been documented to be consistent with those predicted by simple stochastic models.

The most straightforward approximation to the distribution of lineage or taxon sizes is the pure birth process, which entails a certain probability of an individual giving birth or one species giving rise to another per unit time, ignoring death or extinction. In its details, this is clearly an unrealistic model, but it does provide a reasonable first-order approximation for the diversification of higher taxa during their early history, when extinction is infrequent, or in lineages (such as small populations of bacteria growing on a rich medium) where the death rates are very small compared to birth rates.

A basic model (Bailey, 1964; Feller, 1950; Yule, 1924) represents a pure birth process with birth probability λ per unit time. In this model, the instantaneous rate of change in the probability of observing n individuals is given by the differential equation

$$\frac{dp_n}{dt} = \lambda(n-1)p_{n-1} - \lambda np_n. \quad (1.1)$$

Under the initial condition $p_1(0) = 1$ (a single ancestor for the lineage at time zero) and 0 for all $n > 1$, this differential equation has the solution

$$p_n(t) = e^{\lambda t}(1 - e^{\lambda t})^{n-1}. \quad (1.2)$$

More realistically, if there is an intrinsic death rate (or, for species and higher taxa, extinction rate) μ , then one has a time-homogeneous birth and death process defined by the differential equation, e.g., Kendall (1948), Darwin (1954), and Keiding (1975),

$$\begin{aligned} \frac{dp_n}{dt} &= \lambda(n-1)p_{n-1} - (\lambda + \mu)np_n + \mu(n+1)p_{n+1} \quad \text{for } n > 0, \\ \frac{dp_0}{dt} &= \mu p_1. \end{aligned} \quad (1.3)$$

The solutions for $\lambda \neq \mu$ given $p_1(0)=1$ are given by

$$\begin{aligned} p_n(t) &= \frac{(\lambda - \mu)^2 \exp[-(\lambda - \mu)t]}{(\lambda - \mu \exp[-(\lambda - \mu)t])^2} \left(\frac{\lambda - \lambda \exp[-(\lambda - \mu)t]}{\lambda - \mu \exp[-(\lambda - \mu)t]} \right)^{n-1} \quad \text{for } n > 0, \\ p_0(t) &= \frac{\mu - \mu \exp[-(\lambda - \mu)t]}{\lambda - \mu \exp[-(\lambda - \mu)t]}. \end{aligned} \quad (1.4)$$

In the special case where $\lambda = \mu$, the solutions simplify to:

$$\begin{aligned} p_n(t) &= \frac{(\lambda t)^{n-1}}{(\lambda t + 1)^{n+1}} \quad \text{for } n > 0, \\ p_0(t) &= \frac{\lambda t}{\lambda t + 1}. \end{aligned} \quad (1.5)$$

For the purposes of many empirical studies, in which one does not know the precise number of once-extant lineages which are currently extinct (e.g., $n = 0$

individuals in the lineage in a clade at time t , with an indeterminate number in previous time intervals), it is useful to work with the truncated distributions describing the probability densities for strictly positive values of n , i.e.,

$$P_n(t) = \frac{p_n(t)}{1 - p_0(t)} \quad \text{for } n \geq 1$$

which, using the expressions in (1.4), evaluates to

$$P_n(t) = \frac{(\lambda - \mu) \exp[-(\lambda - \mu)t]}{\lambda - \mu \exp[-(\lambda - \mu)t]} \left(\frac{\lambda - \lambda \exp[-(\lambda - \mu)t]}{\lambda - \mu \exp[-(\lambda - \mu)t]} \right)^{n-1} \quad \text{for } n \geq 1, \quad (1.6)$$

while for $\lambda = \mu$,

$$P_n(t) = \frac{(\lambda t)^{n-1}}{(\lambda t + 1)^n} \quad \text{for } n \geq 1. \quad (1.7)$$

Strictly speaking, in order to compare the distributions of family sizes to those predicted in (1.4) or (1.6), or to infer birth and death rates using maximum likelihood estimators, the lineage ages must be known. This requirement becomes particularly problematic when one compares the number of individuals in sublineages, for example the number of species in different genera within a taxonomic family or order. Because the sublineages arose at different times (i.e., obviously, not all genera within a clade originated at the same moment), comparisons of the number of species in a genus or any other subclade must be weighted by the differences in subclade ages.

When the age of a lineage is known, as is the case for phylogenetic trees with branch lengths that have been calibrated using fossil data and “molecular clocks” (Kimura, 1980; Zuckerkandl and Pauling, 1965), the number of individuals or species in each lineage can be compared to the predictions of Eq. (1.4) by applying the appropriate time values. For phylogenies with known branch lengths and resolved split times, Harvey et al. (1994) and Nee et al. (1994, 1995) presented a method whereby the origination and extinction rates λ and μ could be estimated using maximum likelihood methods. A similar approach, not requiring conditioning on the age of the first split, is given by Rannala (1997) and in Felsenstein (2004, pp. 564–569).

Such detailed data is not reliably available for the majority of phylogenetic trees or gene genealogies, and when it is, the confidence intervals on the age estimates can be very wide. To address the problem of rate estimation when molecular clock estimates are absent or unreliable, Reed and Hughes (2002) derived distributions of genera sizes weighted by the time intervals separating generic branching events. Under the assumption that in a clade, subclades such as genera within families arise at a rate ρ , the authors used an exponential approximation for the distribution of the generic ages. This representation is not limited to any taxonomic group, as it applies to any lineage in which constituent sublineages are defined by some unique characters, genetic markers, or in the case of humans, family names.

The time at which a genus within a clade of age τ arises is given by

$$f(t) = \frac{\rho e^{-\rho t}}{1 - e^{-\rho \tau}} \quad (1.8)$$

which, for $\tau \rightarrow \infty$, gives

$$f(t) = \rho e^{-\rho t}. \quad (1.9)$$

Using (1.9) to weight the probability distributions p_n at time t , Reed and Hughes defined

$$q_n(\tau) = \int_0^\tau \rho p_n(t) e^{-\rho t} dt. \quad (1.10)$$

Reed and Hughes derived a generating function and an algorithm for calculation of $q_n(\tau)$ recursively from a limiting case (when $n \rightarrow \infty$). It will be shown that given certain choices of changes in variables, a closed form expression for this integral can be evaluated, and a number of useful approximations to the $q_n(\tau)$ distribution can be made, of which the asymptotic approximation calculated by Reed and Hughes is an important limiting case.

2. Results

2.1. Pure Birth Process

We will begin with the distribution for a pure birth process, which is a reasonable approximation when $\lambda \gg \mu$:

$$q_n(\tau) = \int_0^\tau \rho \exp[-(\lambda + \rho)t] (1 - \exp[-\lambda t])^{n-1} dt. \quad (2.1)$$

Under the assumption that the lineage is ancient relative to the age of any sublineage, we take the limit as $\tau \rightarrow \infty$. Then, applying a change of variables $y = e^{-\lambda t}$, the resulting integral is

$$q_n = \rho \int_0^1 y^{\rho/\lambda} (1 - y)^{n-1} dy \quad (2.2)$$

which, for $n \geq 1$, evaluates to

$$q_n = \text{Beta}(\rho/\lambda + 1, n) = \frac{\Gamma(\frac{\rho}{\lambda} + 1) \Gamma(n)}{\Gamma(n + \frac{\rho}{\lambda} + 1)}. \quad (2.3)$$

Using an asymptotic expansion of the gamma function ratios involving n above (e.g., Exton, 1978; Luke, 1969), we find that

$$q_n = \frac{\Gamma(\rho/\lambda + 1)}{n^{\rho/\lambda + 1}} \left\{ 1 - \frac{(\rho/\lambda)(\rho/\lambda + 1)}{2n} + \frac{(\rho/\lambda + 1)(\rho/\lambda + 2)}{24n^2} \right. \\ \left. \times \left[3 \left(\frac{\rho}{\lambda} + 1 \right)^2 + \frac{\rho}{\lambda} \right] + O(n^{-3}) \right\}.$$

It follows from the above that, as $n \rightarrow \infty$, the probability can be approximated by

$$q_n \approx \Gamma(\rho/\lambda + 1) n^{-\rho/\lambda - 1}. \quad (2.4)$$

2.2. Birth and Death Processes

As in the analyses of Reed and Hughes (2002), we consider three different birth and death models. The first, with $\lambda > \mu$ (birth rate higher than death rate), is perhaps the most biologically significant, because this is the only birth and death process in which a lineage has a non zero probability of persisting indefinitely.

We will evaluate the function Q_n for the truncated solutions to the birth and death process given by Eq. (1.6), because most of the relevant empirical data involves $n > 0$. For convenience we use the parameters $\omega = \lambda - \mu$ and $\theta = \mu/\lambda$, so that when the birth rate is greater than the death rate, $\omega > 0$ and $\theta < 1$. The integral over the truncated distribution can be written as

$$Q_n = \int_0^\infty \frac{\rho\omega \exp[-(\omega + \rho)t]}{1 - \theta \exp[\omega t]} \left(\frac{1 - \exp[-\omega t]}{1 - \theta \exp[-\omega t]} \right)^{n-1} dt.$$

Applying the changes of variable

$$y(t) = \frac{1 - \exp[-\omega t]}{1 - \theta \exp[-\omega t]}, \quad dt = \frac{(1 - \theta)dy}{\omega(1 - y)(1 - \theta y)}$$

and noting that the integral limits are $y(0) = 0$ and $y(\infty) = 1$, we obtain

$$Q_n = \rho(1 - \theta) \int_0^1 y^{n-1} (1 - y)^{\rho/\omega} (1 - \theta y)^{-\rho/\omega-1} dy. \tag{2.5}$$

If we do the same for Eq. (1.4) by integrating over the part of the distribution where $n \geq 1$, an expression of similar form is derived:

$$q_n = \frac{\rho(1 - \theta)}{\omega} \int_0^1 y^{n-1} (1 - y)^{\rho/\omega} (1 - \theta y)^{-\rho/\omega} dy. \tag{2.6}$$

The two integrals given above evaluate to constant multiples of a hypergeometric function (Exton, 1978; Luke, 1969), which has the integral representation

$$\int_0^1 y^{a-1} (1 - y)^{c-a-1} (1 - \theta y)^{-b} dy = \frac{\Gamma(c - a)\Gamma(a)}{\Gamma(c)} {}_2F_1(a, b, c, \theta) \tag{2.7}$$

where ${}_2F_1$ is the Gauss hypergeometric function that is defined by the series

$${}_2F_1(a, b, c, \theta) = \sum_{k=0}^\infty \frac{(a)_k (b)_k}{(c)_k} \theta^k, \tag{2.7a}$$

where

$$(a)_k = a(a + 1) \cdots (a + k - 1) = \frac{\Gamma(a + k)}{\Gamma(a)}$$

(Exton, 1978; Luke, 1969). In Eqs. (2.5) and (2.6), we have $a = n$ and $b = 1 + \rho/\omega$. In the first instance, $c = n + \rho/\omega + 1$, and $c = n + \rho/\omega + 2$ in the second. Using

(2.7) and (2.7a) in (2.5) we obtain

$$Q_n = \rho(1 - \theta)\Gamma\left(1 + \frac{\rho}{\omega}\right) A \sum_{k=0}^{\infty} \left(1 + \frac{\rho}{\omega}\right)_k B \frac{\theta^k}{k!}. \quad (2.8)$$

The coefficients $A = A(\rho, \omega, n)$ and $B = B(\rho, \omega, k, n)$ are given by

$$A(\rho, \omega, n) = \frac{\Gamma(n)}{\Gamma\left(n + \frac{\rho}{\omega} + 1\right)},$$

$$B(\rho, \omega, k, n) = \frac{(n)_k}{(n + \rho/\omega + 1)_k} = \frac{\Gamma(n+k)\Gamma(n + \rho/\omega + 1)}{\Gamma(n)\Gamma(n+k + \rho/\omega + 1)}.$$

It is well known that the ratios of Gamma functions can be expanded in terms of n^{-1} ; see, for example, Luke (1969). The expansion is in terms of Bernoulli polynomials. For similar asymptotic expansions of hypergeometric functions, see also Moschopoulos and Mudholkar (1983). The approximations to follow are straightforward, albeit lengthy and tedious to derive. Details for the derivations of the asymptotic and second-order approximations based on polynomial expansions of Gamma functions have been posted by the authors on the arXiv preprint server (Moschopoulos and Shpak, 2009).

The second-order approximation to the probability function is

$$Q_n = \rho(1 - \theta)\Gamma\left(1 + \frac{\rho}{\omega}\right) n^{-1 - \frac{\rho}{\omega}} \left[1 - \frac{\frac{\rho}{\omega} \left(1 + \frac{\rho}{\omega}\right)}{2n} + \frac{\frac{\rho}{\omega} \left(1 + \frac{\rho}{\omega}\right) \left(3\left(\frac{\rho}{\omega}\right)^2 + \frac{\rho}{\omega}\right)}{24n^2} \right]$$

$$\times \sum_{k=0}^{\infty} \left(1 + \frac{\rho}{\omega}\right)_k \left[1 - \frac{k}{n} \left(1 + \frac{\rho}{\omega}\right) + \frac{\phi(\rho, \omega, k) \theta^k}{n^2 k!} \right] \times (1 + O(n^{-3})), \quad (2.9)$$

where

$$\phi(\rho, \omega, k) = k^2 \left(\frac{\rho^2}{2\omega^2} + \frac{3\rho}{2\omega} + 1 \right) - k \left(\frac{\rho^2}{2\omega^2} \frac{\rho}{2\omega} \right).$$

The terms in the above approximation underscore the fact that the shape of the distribution Q_n is determined by the magnitude of ρ/ω , the ratio of the rates at which new lineages (genera) arise to the net birth rate of individuals or species. This is because in the limit as $\tau \rightarrow \infty$, the absolute magnitudes of the two rate parameters do not contribute to the ratio of lineage number to lineage size. Basically, a small value of ρ/ω implies that over sufficient time, there will be a comparatively low number of lineages rich in individuals, while a large value results in many lineages with relatively few individuals.

It is also remarked that the series terms

$$\sum_{k=0}^{\infty} k \left(1 + \frac{\rho}{\omega}\right)_k \frac{\theta^k}{k!}, \quad \sum_{k=0}^{\infty} k^2 \left(1 + \frac{\rho}{\omega}\right)_k \frac{\theta^k}{k!}$$

have to be computed numerically, just as the hypergeometric function must be estimated using numerical methods. As there is no closed-form expression for Eq. (2.9), the value of these approximations is to show the rate of convergence of the

first- and second-order terms (as functions of n^{-1}) to the exact solution for different birth and death rates.

In contrast, a closed form expression does exist for the asymptotic limit, as $n \rightarrow \infty$. Unlike the terms involving higher powers of k in (2.9), the first series term within the parenthesis converges to

$$\sum_{k=0}^{\infty} \left(1 + \frac{\rho}{\omega}\right)_k \frac{\theta^k}{k!} = (1 - \theta)^{-1-\rho/\omega}.$$

By ignoring all factors of higher order than n^{-1} , we have

$$Q_n \approx \rho(1 - \theta)^{1-\rho/\omega} \Gamma\left(1 + \frac{\rho}{\omega}\right) n^{-\rho/\omega-1} \tag{2.10}$$

as in Reed and Hughes (2002), apart from the constant factor of $1/\omega$ for the truncated distribution. It is noteworthy that in Eq. (2.10) the term θ only appears as a constant factor, while in (2.10) its magnitude determines the rate at which the series term converges. The efficacy of both the asymptotic and n^{-2} approximations will be investigated numerically in the next section, where they are compared to the exact hypergeometric function solutions.

We next direct our attention to the case of a “declining” lineage where $\lambda < \mu$, i.e., when the birth or speciation rate is less than the death or extinction rate. In this scenario, the parameter values satisfy the inequalities $\omega < 0$ and $\theta < 1$. Using the same changes of variables as in the derivation of Eq. (2.6), the integral defining the probability density is

$$Q_n = \rho(1 - \theta) \int_0^{1/\theta} y^{n-1} (1 - y)^{\rho/\omega} (1 - \theta y)^{-\rho/\omega-1} dy. \tag{2.11}$$

With a further change of variables, $z = \theta y$, we again derive an expression with integral limits of 0 and 1,

$$Q_n = \rho(1 - \theta) \theta^{1-n} \int_0^1 z^{n-1} \left(1 - \frac{z}{\theta}\right)^{\rho/\omega} (1 - z)^{-\rho/\omega-1} dz. \tag{2.12}$$

The solution to this integral is a hypergeometric function of a form similar to that of Eq. (2.6). Here, the parameters are $a = n$, $b = -\rho/\omega$, and $c = n - \rho/\omega + 1$. In place of θ in the series expansion, we have $1/\theta$, so that (2.11) ultimately evaluates to

$$Q_n = \rho(1 - \theta) \theta^{1-n} \Gamma\left(-\frac{\rho}{\omega}\right) A \sum_{k=0}^{\infty} \left(-\frac{\rho}{\omega}\right)_k \frac{\theta^k}{k!}. \tag{2.13}$$

Note that because $\omega < 0$, $-\rho/\omega > 0$. For the above distribution, the coefficients A and B are

$$A = \frac{\Gamma(n)}{\Gamma\left(n - \frac{\rho}{\omega}\right)}, \quad B(\rho, \omega, k, n) = \frac{(n)_k}{\left(n - \frac{\rho}{\omega} + 1\right)_k} = \frac{\Gamma(n + k) \Gamma\left(n - \frac{\rho}{\omega} + 1\right)}{\Gamma(n) \Gamma\left(n + k - \frac{\rho}{\omega} + 1\right)}.$$

Using the known asymptotic behavior of gamma function ratios, as in the derivation of (2.9), we find that

$$\begin{aligned}
 Q_n &= \rho(1 - \theta)\theta^{1-n} \Gamma\left(-\frac{\rho}{\omega}\right) n^{\frac{\rho}{\omega}} \\
 &\times \left\{ 1 - \frac{\frac{\rho}{\omega}(\frac{\rho}{\omega} + 1)}{2n} + \frac{(\frac{\rho}{\omega} - 1)(\frac{\rho}{\omega} - 2) \left[3\left(1 + \frac{\rho}{\omega}\right)^2 - \frac{\rho}{\omega} - 1 \right]}{24n^2} \right\} \\
 &\times \sum_{k=0}^{\infty} \left(-\frac{\rho}{\omega}\right)_k \left[1 + \frac{k}{n} \left(\frac{\rho}{\omega} + 1\right) + \frac{\phi(\rho, \omega, k)}{n^2} \left(\frac{\theta^k}{k!}\right) \right] + O(n^{-3}). \quad (2.14)
 \end{aligned}$$

Recall that $\omega < 0$, so that the coefficients $-\rho/\omega$ are positive.

The asymptotic approximation to this distribution, calculated for $n \rightarrow \infty$, is:

$$Q_n \approx \rho\theta^n(1 - \theta)^{1-\rho/\omega} \Gamma\left(1 + \frac{\rho}{\omega}\right) n^{-\rho/\omega-1}. \quad (2.15)$$

The special case where $\lambda = \mu$ (exactly zero net growth rate) is not likely to occur in biological systems, therefore we do not analyze this scenario in detail. In those rare instances where net growth rate is close to zero, the distributions can be approximated by using the equations for the cases where growth rate is positive and negative and taking limits as μ approaches λ .

2.3. A Numerical Assessment of the Approximations

The accuracy of the second order (2.9) and asymptotic approximations (2.10) are assessed by comparing their probability densities to those of the exact solutions (2.8) for a range of parameters values. The examples considered here are the biologically most relevant case where the birth rate is higher than the death rate, such that $\omega > 0$ and $\theta < 1$.

Attention is focused on two parameters: ρ/ω and θ . The first is the ratio of the rate at which new lineages arise to the net birth rate of individuals or species within that lineage. As was discussed above in connection with the distribution (2.9), the shape of the probability density function depends principally on the ratio ρ/ω . When $\rho/\omega \ll 1$, the distribution will be characterized by very few lineages with large numbers of individuals, while as $\rho/\omega \rightarrow 1$, the distribution will be skewed to the left due to the presence of many lineages with few (perhaps only a single) individual. The accuracy of the approximation is examined for different values of this ratio, so that in one instance ρ and ω are of approximately the same magnitude, while in another ρ and ω differ by an order of magnitude.

It is expected from the summation term in Eq. (2.9) that the second-order approximation should be most accurate when $\theta \ll 1$, corresponding to a low birth rate relative to the death rate, because the series converges rapidly when θ is near zero. To compare the accuracy of the approximations as a function of the rate parameters, we will consider cases where of $\theta = 0.01, 0.1, \text{ and } 0.4$. The predictions on the accuracy of the respective approximations are given qualitative support by Figs. 1 and 2, which plot the cumulative density functions

$$\sum_{N=1}^n Q_N$$

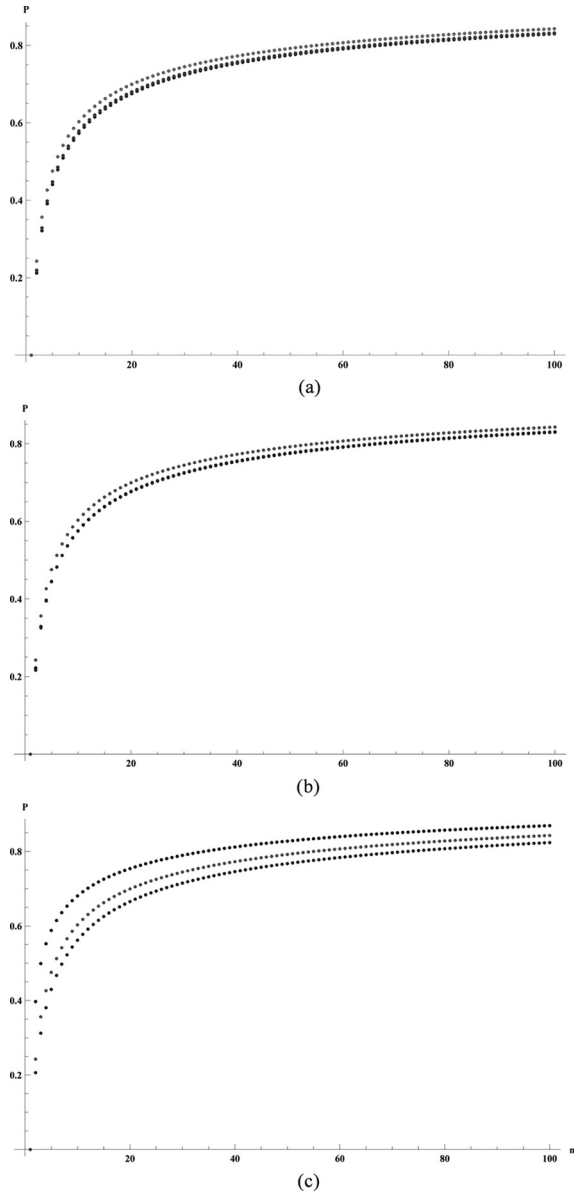


Figure 1. This set of figures shows plots of the cumulative density functions for the exact solution (Eq. (2.8)), the second-order approximation (2.9), and the asymptotic approximation for the case where $\rho = 0.02$ and $\omega = 0.05$. (a) Here $\theta = 0.01$, so that the second-order approximation converges closely to the exact solution (uppermost). The lower curve represents the asymptotic distribution, while the exact solution and second-order approximation are nearly superimposed. (b) For $\theta = 0.1$, the second-order approximation remains close to the exact solution. Note that the asymptotic is independent of θ , so that its divergence from the exact solution is constant in a–c. (c) When $\theta = 0.4$, the second-order approximation (uppermost) deviates significantly from the exact solution (lowermost), even when the sum is taken to the 10^6 . This suggests that first and second-order approximations are only robust when the death rate is sufficiently smaller than the birth rate.

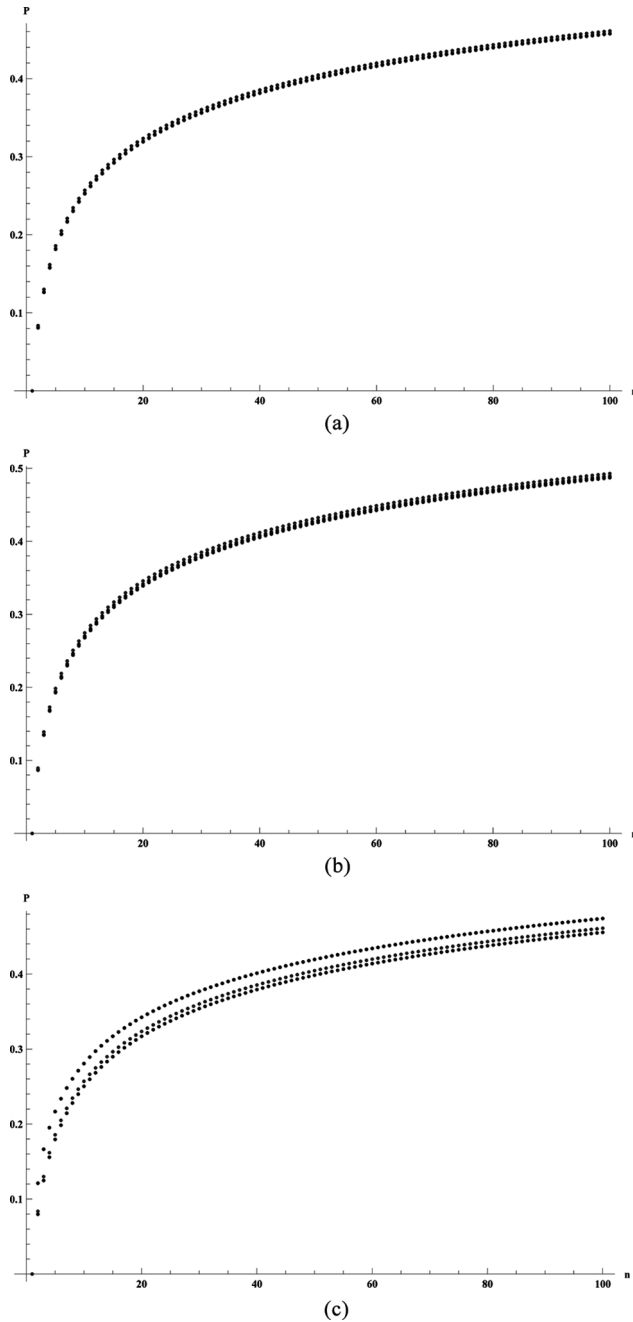


Figure 2. These figures show cumulative density functions $\text{Prob}[N(t) \leq n \mid \frac{\rho}{\omega}, \theta]$ for the same distributions as in Fig. 1, but in this instance $\rho = 0.01$ and $\omega = 0.1$. (a) Again, $\theta = 0.01$, so the second-order approximation is very close. Furthermore, because ρ/ω is fourfold smaller than in Fig. 1, the asymptotic approximation is much closer to the exact distribution. (b) For $\theta = 0.1$, the second-order approximation is still only minimally divergent. The asymptotic remains a robust approximation in this instance, being only dependent on ρ/ω . (c) When $\theta = 0.4$, the same is observed as in Fig. 1(c), i.e., the second order of approximation is the uppermost, exact solution is lowermost graph.

for the exact solution Q_n given by Eq. (2.8), the order n^{-2} approximation of Eq. (2.9), and the asymptotics in Eq. (2.10). In the first set of figures, the rates at which new lineages (e.g., “genera”) are born is of the same order of magnitude as the net birth rate, i.e., $\rho = 0.02$, $\omega = 0.05$, while in the second set, there is a tenfold difference between the birth rates of lineages and individuals, with $\rho = 0.01$ and $\omega = 0.1$.

The relationship between the size of θ and the accuracy of approximation (2.9) can be seen by comparing Figs. 1(a) and 2(a) with 1(c) and 2(c). In Figs. 1(a) and 2(a), $\theta = 0.01$, so that the second order approximation performs better than the asymptotic approximation. The exact solutions and second order approximations start to diverge when $\theta = 0.1$, both in Figs. 1(b) and 2(b). For the case where $\theta = 0.4$, the second order approximation is quite poor, being consistently outperformed by the asymptotic approximation. This is because Eq. (2.10), unlike (2.9), has a closed form expression and does not itself have to be estimated numerically.

As predicted, the asymptotic approximation begins to break down when the net birth rate as ρ , the rate at which new sublineages arise, approaches the value of ω , the birth or speciation rate. This is made apparent by comparing the difference between the exact solutions and the asymptotic approximations in Figs. 1(a–c) (where ρ is of the same magnitude as ω) with those shown Figs. 2(a–c), where ρ is an order of magnitude smaller than ω . The greater deviation between the asymptotic and the exact solutions is due to the fact that when ρ/ω is reasonably large, the coefficients of the n^{-1} and n^{-2} terms cannot be ignored, except when n is very large (of the order of $\sim 10^3$ or greater), so that the second order approximation can match the exact distribution quite closely given parameter values where the asymptotic approximation is inaccurate.

A more precise assessment of the approximation accuracies can be made by comparing cumulative distributions functions. These are given in Table 1 for $n = 10, 100, 200, 500, 1,000$, and $10,000$ are compared to exact and asymptotic values for the same set of birth and death rates that were shown in Figs. 1–2.

For the smaller values of θ , the cumulative probabilities of the second order approximation are very close to the exact distribution, while for $\theta = 0.5$, the asymptotic is better. As was seen in the graphs, the asymptotic cumulative probabilities diverge for ρ of the same order of magnitude as ω . The asymptotic approximation are independent of θ , because this value only appears as a constant factor in Eq. (2.10), while it is a parameter in the hypergeometric function in (2.8) and the approximating sum in (2.9).

Another measurement of the approximation was based on a numerical calculation of percentiles (the calculations were implemented using the *Mathematica* software package, with scripts available from the corresponding author by request) for each distribution, for $p = 0.05$ and $p = 0.01$. The percentile values are shown in Table 2.

As predicted from the previous results, it can be seen that the critical values n^* corresponding to each percentile are very similar for the exact and second order approximate distributions when $\theta \ll 1$. Similarly, the critical values n^* between the exact and second order diverge for larger θ , just as those of the asymptotic and exact distributions diverge for larger values ρ/ω .

Taken together, these examples illustrate the efficacy of the approximations under a range of rate parameter values. It remains to be determined whether the

Table 1

The cumulative probability densities $\text{Prob.}[N(t) \leq n | \frac{\rho}{\omega}, \theta]$ of the exact solution (*E*), the second-order approximation (*S*), and the asymptotic approximation (*A*) are shown for $n = 10, 20, 50, 100, 200, 500, 1,000, 2,000,$ and $10,000$, up to three significant digits. In the first half of the table, the ratio of lineage origination rate to net birth rate $\rho/\omega = 0.25$, while $\rho/\omega = 0.1$ in the second portion. The densities are compared for $\theta = 0.01, 0.1,$ and 0.4 , again showing the breakdown of the second-order approximation for large θ , and of the asymptotic approximation for larger ρ/ω

ρ/ω	θ	$P(10)$	$P(50)$	$P(100)$	$P(500)$	$P(1,000)$	$P(2,000)$	$P(10,000)$	Dist
0.4	0.01	0.580	0.780	0.834	0.915	0.937	0.953	0.977	<i>E</i>
0.4	0.01	0.575	0.777	0.832	0.914	0.936	0.956	0.977	<i>S</i>
0.4	0.01	0.604	0.794	0.845	0.920	0.941	0.956	0.979	<i>A</i>
0.4	0.1	0.577	0.778	0.833	0.914	0.936	0.953	0.978	<i>E</i>
0.4	0.1	0.576	0.777	0.832	0.914	0.936	0.952	0.977	<i>S</i>
0.4	0.1	0.604	0.794	0.844	0.920	0.941	0.956	0.979	<i>A</i>
0.4	0.4	0.563	0.770	0.826	0.911	0.933	0.951	0.976	<i>E</i>
0.4	0.4	0.682	0.829	0.871	0.934	0.950	0.963	0.982	<i>S</i>
0.4	0.4	0.604	0.794	0.845	0.920	0.941	0.956	0.979	<i>A</i>
0.1	0.01	0.271	0.429	0.490	0.616	0.664	0.709	0.803	<i>E</i>
0.1	0.01	0.270	0.428	0.489	0.615	0.664	0.709	0.802	<i>S</i>
0.1	0.01	0.275	0.432	0.493	0.618	0.666	0.711	0.804	<i>A</i>
0.1	0.1	0.270	0.429	0.489	0.615	0.664	0.709	0.802	<i>E</i>
0.1	0.1	0.268	0.425	0.487	0.614	0.662	0.708	0.801	<i>S</i>
0.1	0.1	0.275	0.432	0.493	0.618	0.666	0.711	0.804	<i>A</i>
0.1	0.4	0.268	0.426	0.487	0.614	0.662	0.708	0.802	<i>E</i>
0.1	0.4	0.300	0.448	0.506	0.628	0.674	0.718	0.810	<i>S</i>
0.1	0.4	0.275	0.432	0.493	0.618	0.666	0.711	0.804	<i>A</i>

assumptions made in the derivations are realistic or sufficiently accurate for the biological processes being modeled.

3. Discussion

The results derived in this article give analytical predictions for the number of individuals per lineage in a distribution given a time-homogeneous birth and death process, weighted by a time-homogeneous birth rate for sublineages. An obvious application of these distributions is to the problem that motivated this and similar earlier studies—the number of species within a genus or some higher taxonomic category. With modern techniques of phylogenetic inference, monophyletic subclades of any clade of interest can be identified (together with reasonable bounds on age estimates in many instances) to yield an empirical frequency distribution of subclade sizes. These frequency distributions can then be compared to the probability densities predicted from the exact solutions Q_n .

It is of particular interest to look at “anomalously” species rich subclades and evaluate the probability of their occurrence based on a null model, something which can readily be done by computing the extremal probabilities. For example, if the

Table 2

The critical numbers of individuals n^* corresponding to extremal probability densities $p = 0.05$ and 0.01 are shown for the same parameter values and the same distributions as in Table 1

ρ/ω	θ	$p = 0.05$	$p = 0.01$	Dist
0.4	0.01	1743	49111	<i>E</i>
0.4	0.01	1798	50276	<i>S</i>
0.4	0.01	1491	43576	<i>A</i>
0.4	0.1	1778	49853	<i>E</i>
0.4	0.1	1801	50358	<i>S</i>
0.4	0.1	1491	43576	<i>A</i>
0.4	0.4	1949	53439	<i>E</i>
0.4	0.4	980	31356	<i>S</i>
0.4	0.4	1491	43576	<i>A</i>
0.1	0.01	251193	746770	<i>E</i>
0.1	0.01	251582	747932	<i>S</i>
0.1	0.01	249229	745463	<i>A</i>
0.1	0.1	251470	746954	<i>E</i>
0.1	0.1	252951	747932	<i>S</i>
0.1	0.1	249229	745463	<i>A</i>
0.1	0.4	252810	747839	<i>E</i>
0.1	0.4	241228	740048	<i>S</i>
0.1	0.4	249229	745463	<i>A</i>

average rate at which new genera arise in an extant clade is 0.02 per million years, the net speciation rate $\omega = 0.05$, and the ratio of extinction to speciation rate is 0.1, then it can be seen from the appropriate entry in Table 1 the probability of encountering a genus with over 10,000 species is less than 0.025 according to the distributions derived under the assumptions of time homogeneity. If such unusually large genera are encountered, that may indicate that the birth-death process driving the distribution is in fact not time-homogeneous.

Another application of these results to phylogenetic data is the estimation of speciation and extinction rates from taxon size distributions. The most straightforward, and by far the most accurate, means of estimating these parameters is from direct paleontological data on origination and extinction rates, which is rarely available at the level of detail required for likelihood based estimates. The approach of Harvey and colleagues (e.g., Harvey et al., 1994; Nee et al., 1994) does not require fossil data, relying instead upon molecular clock estimates of subclade ages. In addition to the fact that the confidence intervals on these estimators tend to be very large (Paradis, 2004), there is the fact that reliable clock data is not available for many taxa. Consequently, the distributions in Eqs. (2.8)–(2.10) can be used to give numerical estimates of the parameters, much as Reed and Hughes (2002) obtained recursively and in asymptotic approximations.

The underlying processes for which the distributions were derived are not limited to phylogenetic data on species abundance. Indeed, any process generating new entities defined as lineages or families, that are themselves composed of

individual entities undergoing birth and death, can be analyzed using the models presented in this paper. For example, gene genealogies within populations and species often have constituent lineages that are defined by a set of point mutations. This is of course the foundation of coalescent based approaches to population genetics, whereby the history of lineages can be inferred from sequence data (e.g., Hein et al., 2005; Hudson, 1991; Kingman, 1982; Wakeley, 2008). The diversity of genotypes and lineages is produced by a branching process where the birth and death rates can be parameterized by λ and μ , while the mutation rate defining new lineages in infinite sites or infinite alleles models can be used as a measure of ρ . In this way, haplotype distributions in population genetics can be analyzed in the same way as the taxonomic data described above, allowing one to determine whether the disproportionate representation of certain alleles or haplotypes is consistent with a random branching process, or whether the haplotype distribution must be explained by processes other than mutation and genetic drift.

The correspondence between the data and the analytical results in all of these cases will to a large part be limited by a number of simplifying assumptions made in the derivations. In particular, the distributions Q_n were derived under the assumption that the time elapsed since the initiation of the process is very great, essentially infinite. Therefore, for comparatively young lineages (young with respect to the birth and death rates), one would expect the approximations to break down. These issues will be investigated using individual-based simulations and analysis of biological data in a forthcoming article.

Acknowledgments

The authors thank an anonymous reviewer for comments and suggested corrections. Panagis Moschopoulos was supported by RCMI/NIH grant 5G12 RR008124 from the National Institutes of Health to the Border Biomedical Research Center (BBRC) at the University of Texas at El Paso (UTEP), and Max Shpak was supported by funding for research from the University of Texas at El Paso.

References

- Anderson, S. (1974). Patterns of faunal evolution. *Quarter. Rev. Biol.* 49:311–332.
- Bailey, N. T. J. (1964). *The Elements of Stochastic Processes with Applications to the Natural Sciences*. New York: John Wiley and Sons.
- Burlando, B. (1990). The fractal dimension of taxonomic systems. *J. Theoret. Biol.* 146:99–114.
- Darwin, J. H. (1954). The behavior of an estimator for a simple birth and death process. *Biometrika* 43:23–31.
- Exton, H. (1978). *Handbook of Hypergeometric Integrals*. New York: John Wiley and Sons.
- Feller, W. (1950). *Probability Theory and Its Applications*. New York: John Wiley and Sons.
- Felsenstein, J. (2004). *Inferring Phylogenies*. Sunderland: Sinauer Associates.
- Harvey, P. H., May, R. M., Nee, S. (1994). Phylogenies without fossils. *Evolution* 48:523–529.
- Hein, J., Schierup, M. H., Wiuf, C. (2005). *Gene Genealogies, Variation, and Evolution: A Primer in Coalescent Theory*. Oxford: Oxford University Press.
- Hudson, R. R. (1991). Gene genealogies and the coalescent process. *Oxford Surv. Evol. Biol.* 7:1–49.
- Keiding, N. (1975). Maximum likelihood estimation in the birth and death process. *Ann. Statist.* 3:363–372.

- Kendall, D. G. (1948). On the generalized birth and death process. *Ann. Mathemat. Statist.* 19:1–15.
- Kimura, M. (1980). *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge University Press.
- Kingman, J. F. C. (1982). The coalescent. *Stochastic Process. Applic.* 13:235–248.
- Luke, Y. L. (1969). *The Special Functions and Their Approximations*. New York: Academic Press.
- Moschopoulos, P. G., Mudholkar, G. S. (1983). A likelihood ratio based normal approximation for the non-null distribution of the multiple correlation coefficient. *Commun. Statist. Simul. Computat.* 12:355–371.
- Moschopoulos, P. G., Shpak, M. (2009). Taxon size distribution in a time-homogeneous birth and death process. arXiv:0901.1066v1
- Nee, S., May, R. M., Harvey, P. H. (1994). The reconstructed evolutionary process. *Phil. Trans. Roy. Soc. London B* 344:305–311.
- Nee, S., Holmes, E. C., May, R. M., Harvey, P. H. (1995). Estimating extinction rates from molecular phylogenies. In: Lawton, J. H., May, R. M., eds. *Extinction Rates*. Oxford: Oxford University Press, pp. 164–182.
- Paradis, E. (2004). Can extinction rates be estimated without fossils? *J. Theoret. Biol.* 229:1–30.
- Rannala, B. (1997). Gene genealogy in a population of variable size. *Heredity* 78:417–423.
- Reed, W. J., Hughes, B. D. (2002). On the size distribution of live genera. *J. Theoret. Biol.* 217:125–135.
- Wakeley, J. (2008). *Coalescent Theory: An Introduction*. Greenwood Village: Roberts Publishing.
- Watson, H. W., Galton, F. (1875). On the probability of extinction of families. *J. Anthropol. Instit. Great Brit.* 4:138–144.
- Yule, G. U. (1924). A mathematical theory of evolution based on the conclusions of Dr. J. C. Willis. *F.R.S. Phil. Trans. Roy. Soc. London B* 213:21–87.
- Zuckermandl, E., Pauling, L. (1965). Molecules as documents of evolutionary history. *J. Theoret. Biol.* 8:357–366.