An Asymptotic Result in a Pure Birth Stochastic Process

Panagis G. Moschopoulos

Univ. of Texas at El Paso
Corresponding author email: pmoschopoulos@utep.edu

Abstract: The ${}_2F_1$ hypergeometric function appears in a wide variety of applications in probability and statistics. This article is motivated from an important application that arises in stochastic processes. A birth and death process that starts with a single species and birth rate larger than the death rate may lead to trees of species or genera. Under standard assumptions on the process, the probability of the size of a genera is a discrete distribution that can easily be computed at time t from the ori-gin. However, when the process leads to several sub-lineages of trees that originated at different times from the origin, then the distribution of the size of the population commonly referred to as taxon size distribution of the genus size must be weighted by the different times that the genera originate. It was shown in [1], that under the exponential distribution for time t the taxon distribution may be ex-pressed as an ${}_2F_1(x, b; x + b; \theta)$ hypergeometric function. This form has not been exploited in the literature. In this article we consider an asymptotic expansion of this ${}_2F_1$ for large x in terms of generalized Bernoulli polynomials that can be computed up to any degree of accuracy.

Keywords: birth-death process, taxon size, hypergeometric function, generalized Bernoulli polynomials, asymptotic expansion.

1. Introduction

In a recent article [1], the researchers considered the number of individuals in a genetic lineage (family size or genus size) under a time homogeneous process that arises from a birth and death stochastic process [8]. The latter is a well-known process but we state the fundamentals here for clarity [9]. Let λ be the birth rate and be the death rate. The case $\mu = 0$ is the pure birth process and it was essentially introduced in [4], see also [5]. If we denote by $P_n(t)$ the probability that there are n individuals at time t, then the instantaneous rate of change in the probability $P_n(t)$at time t is given by the di erential equation

$$\frac{dP_n(t)}{dt} = \lambda(n-1)P_{n-1}(t) - \lambda_n P_n(t) \quad (1.1)$$

Under the initial condition $P_1(0) = 1$, i.e. a lineage starts with a single ancestor, and 0 for $n > 1$, (1.1) has the solution

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

If there is an intrinsic death rate $\mu(> 0)$, then the differential equation for the birth and death process is

$$\frac{dP_n(t)}{dt} = \lambda(n-1)P_{n-1}(t) - (\lambda + \mu)nP_n(t) + \mu(n+1)P_{n+1}(t), \quad (1.3)$$

for $n \geq 1$, and $P_1(0) = 1$, $\frac{dP_0(t)}{dt} = \mu P_1(t)$.

Under these conditions, the solution to (1.3) is given by

$$P_n(t) = \frac{(\lambda-\mu)^2 e^{-(\lambda-\mu)t}}{[\lambda - \mu e^{-(\lambda-\mu)t}]^2}\left(\frac{\lambda - \lambda e^{-(\lambda-\mu)t}}{\lambda - \mu e^{-(\lambda-\mu)t}}\right)^{n-1} \quad (1.4)$$

$$P_0(t) = \frac{\mu - \mu e^{-(\lambda-\mu)t}}{\lambda - \mu e^{-(\lambda-\mu)t}} \quad (1.5)$$

Where

$$w = \lambda - \mu\,, \theta = \frac{\mu}{\lambda}.$$

In the following we assume $\lambda > \mu$ so that $\omega > 0$ and $\theta < 1$. The case of $\lambda < \mu$ is di erent and is not handled here as is of interest only in extinct populations. Now, if the time t is known, then (1.6) completely de nes the process. However, the time t of origination of some clades (sub-lineages) is usually un-known. This is the case, for example, when we have species in different genera in a taxonomic family. Since not all genera originate at the same time, the total number of species in a genus must be weighted by the different ages of the sub-lineages. Hence, to compute the unconditional distribution of the genus size N, the probability $P_n(t)$ in (1.6) must be weighted by the time distribution.

The exponential distribution has been established in the literature as a reasonably good approximation to the real distribution of times at which the different genera originate, see reasoning in [6]. If the rate that genera originate is $1/\rho$ then (1.6) must be weighted by

$$f(t) = \rho e^{-\rho t}, \rho > 0, t > 0$$

and be integrated over the time t. This leads to the following expression for the unconditional distribution of the total size N of the genus:

$$P(N = n) = \int_0^\infty \frac{\rho\omega e^{-(\omega+\rho)t}}{1 - \theta e^{-\omega t}}\left(\frac{1 - e^{-\omega t}}{1 - \theta e^{-\omega t}}\right)^{n-1} dt. \quad (1.7)$$

2. The P (N = n) as a Hypergeometric Function

In (1.7) we let:

$$e^{-\omega t} = \frac{1-y}{1-\theta y}, 0 < y < 1 \quad (2.1)$$

which gives:

$$dt = \frac{(1-\theta)dy}{\omega(1-y)(1-\theta y)}, \quad (2.2)$$

$$e^{-\rho t} = (e^{-\omega t})^{\frac{\rho}{\omega}} = \frac{(1-y)^{\frac{\rho}{\omega}}}{(1-\theta y)^{\frac{\rho}{\omega}}}. \quad (2.3)$$

Substitution to (1.7) leads to the the following:

$$P(N=n) = q_n = \rho \int_0^1 y^{n-1}(1-y)^{\frac{\rho}{\omega}} = (1-\theta y)^{-\frac{\rho}{\omega-1}} dy \quad (2.4)$$

The expression above is clearly part of the hypergeometric function ${}_2F_1$(a; b; c;) de ned by (integral representation, see for example [7]:

$${}_2F_1(a,b;c;\theta) = \frac{\Gamma(c)}{\Gamma(a)\Gamma(c-a)} \rho \int_0^1 y^{a-1}(1-y)^{c-a-1}(1-\theta y)^{-b} dy \quad (2.5)$$

Recall that we assumed $0 < \theta < 1$ so that the ${}_2F_1$ is convergent. (2.4) and (2.5) we have:

$$a = n, b = \rho/\omega + 1 > 0, c = n + b \quad (2.6)$$

and (2.5) becomes:

$$q_n = \rho \frac{\Gamma(n)\Gamma(b)}{\Gamma(n+b)} \times {}_2F_1(n,b;n+b;\theta). \quad (2.7)$$

3. Asymptotic Expansion of ${}_2F_1(n,b;n+b;\theta)$

The ${}_2F_1$ function is also de ned by the following series (which can also be read-ily obtained from (2.4) by expanding $(1-\theta y)^{-b}$.

$${}_2F_1(n,b;n+b;\theta) = \sum_{k=0}^{\infty} \frac{(n)_k (b)_k \theta^k}{(n+b)_k k!}$$

$$= \frac{\Gamma(n+b)}{\Gamma(n)} \sum_{k=0}^{\infty} (b)_k \frac{\Gamma(n+k)}{\Gamma(n+b+k)} \frac{\theta^k}{k!} \quad (3.1)$$

where $(b)_k = b(b+1)\dots(b+k-1), (b)_0 = 1.$ When $n \to \infty$ with the help of Stirning's formula we see that

$$\frac{\Gamma(n+b)}{\Gamma(n)} \frac{\Gamma(n+k)}{\Gamma(n+b+k)} \to 1$$

and hence the right side of (3.1) reduces to

$$\sum_{k=0}^{\infty} (b)_k \frac{\theta^k}{k!} = (1-\theta)^{-b}$$

for $|\theta| < 1$ and this is the asymptotic result as $n \to \infty$. But successive approximations can be obtained if the gamma rations are expended by using the asymptotic expansions in terms of Bernoulli polynomials.

Now, an asymptotic expansion of the gamma ratios involved in (3.1) can be obtained for large n in terms of generalized Bernoulli polynomials, see for example [8], the result is from [9].

$$\frac{\Gamma(z+a)}{\Gamma(z+b)} = z^{a-b} \sum_{j=0}^{\infty} (-1)^j B_j^{(a-b+1)}(a)(b-a)_j z^{-j}/j! \quad (3.2)$$

The expression is valid for z complex but in our case z = n, a positive integer, normally large since it is the population size. Also required is that b a is bounded. Here $b - a = 1 + \rho/\omega$ and we have assumed that $\omega = \lambda - \mu > 0$ and of course we have $\rho > 0$. $B_j^{(a-b+1)}(a)$ are the generalized Bernoulli polynomials de ned by the generating formula:

$$\frac{t^a e^{xt}}{(e^t-1)^a} = \sum_{k=0}^{\infty} \frac{t^k}{k!} B_k^a(x), |t| < 2\pi \quad (3.3)$$

When $a = 1$ we have $B_k^{(1)}(x) = B_k(x)$, the Bernoullia polynomials, and when $a = 1$ and $x = 0$ we have the Bernoulli numbers $B_k(0) = B_k$. The first six generalized Bernoulli polynomials are given below.

$$B_0^{(a)}(x) = 1, \; B_1^{(a)}(x) = x - \frac{a}{2},$$

$$B_2^{(a)}(x) = x^2 - ax + \frac{a(3a-1)}{12},$$

$$B_3^{(a)}(x) = x^3 - \frac{3a}{2}x^2 + \frac{a(3a-1)}{4}x - \frac{a^2(a-1)}{8},$$

$$B_4^{(a)}(x) = x^4 - 2ax^3 + \frac{a(3a-1)}{2}x^2 - \frac{a^2(a-1)}{2}(x) + \frac{a}{240}(15a^3 - 30a^2 + 5a + 2),$$

$$B_5^{(a)}(x) = x^5 - \frac{50}{2}x^4 + \frac{5a(3a-1)}{6}x^3 - \frac{5a^2(a-1)}{4}x^2 + \frac{a(15a^3-30a^2+5a+2)}{48}x - \frac{a^2(a-1)(3a^2-7a-2)}{96},$$

$$B_6^{(a)}(x) = x^6 - 3ax^5 + \frac{5a(3a-1)}{4}x^4 - \frac{5a^2(a-1)}{2}x^3 + \frac{a(15a^3-30a^2+5a+2)}{16}x^2 - \frac{a^2(a-1)(3a^2-7a-2)}{16}x + \frac{a(63a^5-315a^4+315a^3+91a^2-42a-16)}{4032}$$

Using (3.2) the gamma ratios in (3.1) are expressed as follows:

$$\frac{\Gamma(n+b)}{\Gamma(n)} = n^b \sum_{j=0}^{\infty} (-1)^j B_j^{(b+1)}(b)(-b)_j n^{-j}/j! \quad (3.4)$$

$$\frac{\Gamma(n+k)}{\Gamma(n+b+k)} = n^{-b} \sum_{j=0}^{\infty} (-1)^j B_j^{(-b+1)}(k)(b)_j n^{-j}/j! \quad (3.5)$$

Substituting the above to (3.1) we obtain:

$${}_2F_1(n,b;n+b;\theta) =$$

$$= n^b \sum_{j=0}^{\infty} (-1)^j B_j^{(b+1)}(b)(-b)_j n^{-j}/j!$$

$$\times \left[\sum_{k=0}^{\infty} (b)_k n^{-b} \sum_{i=0}^{\infty} (-1)^i B_i^{(-b+1)}(k)(b)_i n^{-i}/i! \; \theta^k/k!\right]$$

$$= \left(\sum_{j=0}^{\infty} C_j n^{-j}\right) \times \sum_{k=0}^{\infty} (b)_k \left(\sum_{i=0}^{\infty} A_i(k) n^{-i}\right) \theta^k/k! \quad (3.6)$$

Where

$$C_j = (-1)^j B_j^{(b+1)}(b)(-b)_j/j!, \quad j = 0,1,2,\dots \quad (3.7)$$

$$A_i(k) = (-1)^i B_i^{(-b+1)}(k)(b)_i/i! \; \theta^k/k! \quad i = 0,1,2,\dots; k = 0,1,2,\dots \quad (3.8)$$

Further, letting

$$D_j = \sum_{k=0}^{\infty} (b)_k A_j(k), \ \ j = 0,1,2, \ldots, \tag{3.9}$$

We obtain

$$ {}_2\mathrm{F}_1(n, b; n+b; \theta) = \sum_{j=0}^{\infty} C_j n^{-j} \sum_{j=0}^{\infty} D_j n^{-j}$$

$$= \sum_{j=0}^{\infty} E_j n^{-j} \ , \tag{3.10}$$

Where

$$E_j = \sum_{i=1}^{j} C_i D_{j-i}, \ \ j = 0,1,2, \ldots, \tag{3.11}$$

We summarize the result as a Theorem.

Theorem 1. For large n,b bounded, and $|\theta| < 1$ the ${}_2\mathrm{F}_1(n, b; n+b; \theta)$ is expanded in powers of n^{-1} as

$$ {}_2\mathrm{F}_1(n, b; n+b; \theta) = \sum_{j=0}^{\infty} E_j n^{-j} \ ,$$

Where

$$E_j = \sum_{i=1}^{j} C_i D_{j-i}, \ \ j = 0,1,2, \ldots,$$

$$C_j = (-1)^j B_j^{(b+1)}(b)(-b)_j / j!, \qquad j = 0,1,2, \ldots$$

$$D_j = \sum_{k=0}^{\infty} (b)_k A_j(k), \ \ j = 0,1,2, \ldots,$$

$A_j(k) = (-1)^j B_j^{(-b+1)}(k)(b)_j / j! \ \theta^k / k!$, $j = 0,1,2, \ldots$; $k = 0,1,2, \ldots$. □

It follows from the Theorem that the asymptotic $(n \to \infty)$ term of ${}_2\mathrm{F}_1(n, b; n+b; \theta)$ is just E_0 and

$$E_0 = C_0 D_0 = A_j(k)$$

$$= B_0^{(b+1)}(b)(-b)_0 (b)_0 \sum_{k=0}^{\infty} (b)_k B_0^{(-b+1)}(k) \, \theta^k / k!$$

$$= \sum_{k=0}^{\infty} (b)_k \, \theta^k / k! = (1-\theta)^{-b} = D_0, \ \ C_0 = 1 \tag{3.12}$$

We now return to the distribution q_n given in (2.6). Note that the gamma ratio $\Gamma(n)/\Gamma(n+b)$ cancels out with its inverse inside the series, and q_n reduces to a single series involving the D_j

$$q_n = \rho \Gamma(b) n^{-b} \sum_{k=0}^{\infty} (b)_k \sum_{j=0}^{\infty} (-1)^j C_j$$

$$= (-1)^j B_j^{(-b+1)}(k)(b)_j n^{-j} / j! \, \theta^k$$

$$/ k!. \tag{3.13}$$

$$= \rho \Gamma(b) n^{-b} \sum_{j=0}^{\infty} n^{-j} \, D_j$$

$$= \rho \Gamma(\rho/\omega + 1) n^{-\rho/\omega - 1} (1-\theta)^{-\rho/\omega - 1} (1 + O(n^{-1}))$$

The last equation confirms the asymptotic result obtained in [1, 6]. However, (3.13) can be used to any degree of accuracy. Further simplification of the 'large' n distribution or the complete series does not appear feasible.

Acknowledgements

This work was supported by Grant G12MD007592 from the National Institutes on Minority Health and Health Disparities (NIMHD), a component of the National Institutes of Health (NIH).

References

[1] P.G. Moschopoulos, M. Shpak. *The distribution of family sizes under a time homogeneous birth and Death process*, Communications in Statistics-Theory and Methods, vol. 39, no. 10, pp. 1761-1775, 2010.

[2] Norman T.J. Bailey, *The elements of stochastic processes with applications to the natural science*, John Wiley and Sons, New York, vol. 25, 1964.

[3] D.G. Kendall, *On the generalized birth and death process*, Annals of Mathematical Statistics, vol. 19, pp. 1-15, 1948.

[4] G.U. Yul, *A mathematical theory of evolution based on the conclusions of Dr. J.C. Willis*, F.R.S. Philosophical Transactions of the Royal Society of London B. 213, pp. 21-87, 1925.

[5] W. Feller, *Probability theory and its applications*, John Wiley and Sons, New York, 1950.

[6] W.J. Reed, B.D. Hughes, *On the size distribution of live genera*, Journal of Theoretical Biology vol. 217, no. 1, pp. 125-135, 2002.

[7] A.M. Mathai, H.J. Haubold, *Special Functions for Applied Sci-entists*. Springer, N.Y., vol. 4, 2008.

[8] Y.L. Luke, *The special functions and their approximations*, Academic Press, New York., Vol. 53, 1969.

[9] F.G. Tricomi, Erdely, *The asymptotic expansion of a ratio of gamma functions*, Paci c J. Math., vol. 1, no. 1, pp. 133-142, 1951.