A distribution function from population genetics statistics using Stirling numbers of the first kind: Asymptotics, inversion and numerical evaluation

Swaine L. Chen[*]   Nico M. Temme[†]

November 23, 2021

Abstract

Stirling numbers of the first kind are common in number theory and combinatorics; through Ewen's sampling formula, these numbers enter into the calculation of several population genetics statistics, such as Fu's F_s. In previous papers we have considered an asymptotic estimator for a finite sum of Stirling numbers, which enables rapid and accurate calculation of Fu's F_s. These sums can also be viewed as a cumulative distribution function; this formulation leads directly to an inversion problem, where, given a value for Fu's F_s, the goal is to solve for one of the input parameters. We solve this inversion using Newton iteration for small parameters. For large parameters we need to extend the earlier obtained asymptotic results to handle the inversion problem asymptotically. Numerical experiments are given to show the efficiency of both solving the inversion problem and the expanded estimator for the statistical quantities.

Keywords Stirling numbers of the first kind; Asymptotic analysis; Population genetics statistics; Evolutionary inference from sequence alignments; Numerical algorithms; Cumulative distribution function.

1   Introduction

In recent papers [1] and [2] we have discussed the sum

$$S'_{n,m}(\theta) = \frac{1}{(\theta)_n} \sum_{k=m}^{n} (-1)^{n-k} S_n^{(k)} \theta^k, \quad \theta > 0, \qquad (1.1)$$

[*]Infectious Diseases Translational Research Programme and Department of Medicine, Division of Infectious Diseases, Yong Loo Lin School of Medicine, National University of Singapore, Singapore 119228, Singapore & Laboratory of Bacterial Genomics, Genome Institute of Singapore, Singapore 138672, Singapore. Email: slchen@gis.a-star.edu.sg

[†]IAA, 1825 BD 25, Alkmaar, The Netherlands. Former address: Centrum Wiskunde & Informatica (CWI), Science Park 123, 1098 XG Amsterdam, The Netherlands. Email: nico.temme@cwi.nl

1

where $S_n^{(k)}$ are the Stirling numbers of the first kind defined by

$$(\theta)_n = \sum_{k=0}^{n}(-1)^{n-k}S_n^{(k)}\theta^k, \qquad (1.2)$$

and $(\theta)_n$ is the Pochhammer symbol, defined by

$$(\theta)_0 = 1, \quad (\theta)_n = \theta(\theta+1)\cdots(\theta+n-1) = \frac{\Gamma(\theta+n)}{\Gamma(\theta)}. \qquad (1.3)$$

The quantity $S'_{n,m}(\theta)$ (and related quantities) is used in the calculation of several population genetics statistics. One such statistic is Fu's F_s,

$$F_s = \ln\frac{S'_{n,m}(\theta)}{1-S'_{n,m}(\theta)}, \qquad (1.4)$$

which was shown to be capable of identifying the genetic changes responsible for the increased fitness of a recently expanded clone of Campylobacter jejuni that is causing an epidemic of abortion in livestock [10]. We used asymptotic approxima- tions of the Stirling numbers derived in [8] to compute $S'_{n,m}(\theta)$ [1]. Subsequently, we transformed the sum into a contour integral in the complex plane, and we gave a first-order approximation of this integral for large n, with $0 < m < n$ and $\theta > 0$ [2].

Note that, because of (1.2) the sum (1.1) satisfies $0 \leq S'_{n,m}(\theta) \leq 1$; see Figure 1. In fact $S'_{n,m}(\theta)$ can be viewed as a cumulative distribution function frequently used in the derivation of several population genetics statistics, which in turn are useful for testing evolutionary hypotheses directly from DNA sequences.

In the earlier paper [2] we have derived a new integral representation of $S'_{n,m}(\theta)$ and we have given a first-order asymptotic approximation in terms of an incomplete beta function. With these results the algorithm given in a first attempt [1] could be considerably improved in efficiency and speed.

In the present paper the main interest is the inversion problem to find θ from the equation $S'_{n,m}(\theta) = s$ with given $s \in (0,1)$, n and m. For small or intermediate values of n we use a Newton iteration scheme, whereas for large n we derive more details on the earlier derived asymptotic representation of $S'_{n,m}(\theta)$ to develop an asymptotic expansion of the wanted θ. Numerical experiments are given to show the efficiency of the asymptotic expansion, of the Newton iterations, and the asymptotic inversion problem.

2　A few details on the Stirling numbers

For a concise overview of properties, of these Stirling numbers, with a summary of their uniform approximations, see [4, §11.3].

In Figure 1 we show two graphs of $S'_{n,m}(\theta)$, on the left left a point plot for $0 \leq m \leq n$, with fixed $\theta = 50$ and $n = 100$, and on the right a smooth sigmoid curve for $0 \leq \theta \leq n$ with fixed $m = 50$ and $n = 100$.
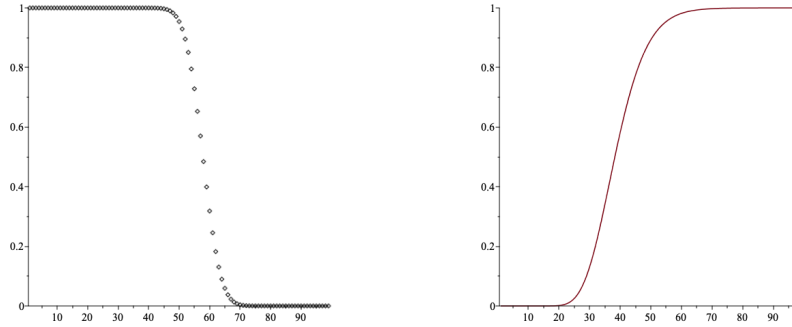
Figure 1: **Left:** The function $S'_{n,m}(\theta)$ for $m = 0, 1, 2, \ldots, 100$; $\theta = 50$ and $n = 100$. **Right:** The function $S'_{n,m}(\theta)$ for $\theta \in [0, 100]$; $m = 50$ and $n = 100$.

A key representations for the considered asymptotic problem is the Cauchy-type integral in the complex plane

$$(-1)^{n-k} S_n^{(k)} = \frac{1}{2\pi i} \int_\mathcal{C} \frac{(z)_n}{z^{k+1}} \, dz, \tag{2.1}$$

which follows from (1.2). Here $\mathcal{C}$ is a contour around the origin.

Special values are

$$S_n^{(n)} = 1 \; (n \geq 0), \quad S_n^{(0)} = 0 \; (n \geq 1), \quad S_n^{(1)} = (-1)^{n-1}(n-1)! \; (n \geq 1), \tag{2.2}$$

and there is a recurrence relation:

$$S_{n+1}^{(k)} = S_n^{(k-1)} - n S_n^{(k)}. \tag{2.3}$$

For the sums $S'_{n,m}(\theta)$ we have a new similar result.

Theorem 2.1. *The sums $S'_{n,m}(\theta)$ satisfy for $n = 2, 3, 4, \ldots$ the recursion*

$$(\theta + n) S'_{n+1,m}(\theta) = n S'_{n,m}(\theta) + \theta S'_{n,m-1}(\theta), \quad 1 \leq m \leq n,$$

$$S'_{n+1,n+1}(\theta) = \frac{\theta^{n+1}}{(\theta)_{n+1}}, \tag{2.4}$$

with initial values

$$S'_{0,0}(\theta) = 1, \quad S'_{1,0}(\theta) = 1, \quad S'_{1,1}(\theta) = 1,$$

$$S'_{2,0}(\theta) = 1, \quad S'_{2,1}(\theta) = 1, \quad S'_{2,2}(\theta) = \frac{\theta}{\theta + 1}. \tag{2.5}$$

Proof. The proof simply follows from using the relation in (2.3). We also need $S_n^{(n)} = 1$ and $(\theta)_{n+1} = (\theta + n)(\theta)_n$. $\qquad\square$

3

Observe that there are no Stirling numbers in the recursion in (2.4), apart from those needed to compute the starting values in (2.5). This gives a very simple method to compute $S'_{n,m}(\theta)$ for small or intermediate values of n. For large values of n we prefer asymptotic representations.

The Stirling numbers are very large for large values of n and $m \ll n$, see the value of $S_n^{(1)}$ in (2.2). This makes straightforward evaluation of the sum in (1.1) sensitive to overflow. This problem does not happen for the recursion in (2.4), because $0 \le S'_{n,m}(\theta) \le 1$.

In the initial values we see that $S'_{n,0}(\theta) = S'_{n,1}(\theta) = 1$ for $n = 0, 1$. More generally this follows from $S_n^{(0)} = 0$ if $n \ge 1$.

In the computation of cumulative distribution functions, like the classical gamma and beta cases, it is essential to consider the complementary relations. In the present case we use the complementary sum

$$T'_{n,m}(\theta) = 1 - S'_{n,m}(\theta) = \frac{1}{(\theta)_n} \sum_{k=0}^{m-1} (-1)^{n-k} S_n^{(k)} \theta^k. \qquad (2.6)$$

To avoid numerical cancellation when using the complementary relation, we should compute first the *primary* function, that is, $\min(S'_{n,m}(\theta), T'_{n,m}(\theta))$, and the other one from the complementary relation. The functions $T'_{n,m}(\theta)$ satisfy the same recursion as $S'_{n,m}(\theta)$ in (2.4), of course with different starting values.

As mentioned above, Fu's F_s in (1.4) is of interest for population genetics applications [3]. From (2.6), we also have a complementary equation for Fu's F_s,

$$F_s = \ln \frac{1 - T'_{n,m}(\theta)}{T'_{n,m}(\theta)}. \qquad (2.7)$$

Fu's F_s ranges from $-\infty$ to $+\infty$ as θ runs through the interval $(0, \infty)$ and it vanishes when θ equals its *transition value* θ_t for which value we have

$$S'_{n,m}(\theta_t) = T'_{n,m}(\theta_t) = \tfrac{1}{2}. \qquad (2.8)$$

Both representations of F_s are needed in numerical computations, because when $S'_{n,m}(\theta)$ is close to 1, the form with $T'_{n,m}(\theta)$ in (2.7) gives a more reliable computational representation.

3 Summary of earlier results

Because it is more convenient to work with $S'_{n+1,m+1}(\theta)$ we proceed with

$$\begin{aligned} S_{n+1,m+1}(\theta) &= \frac{1}{(\theta+1)_n} \sum_{k=m}^{n} (-1)^{n-k} S_{n+1}^{(k+1)} \theta^k, \\ T_{n+1,m+1}(\theta) &= \frac{1}{(\theta+1)_n} \sum_{k=0}^{m-1} (-1)^{n-k} S_{n+1}^{(k+1)} \theta^k. \end{aligned} \qquad (3.1)$$

The following result of our paper [2] is crucial for deriving asymptotic expansions of the statistical quantities.

Theorem 3.1. *Let $\mathcal{C}_\rho$ be a circle at the origin of the complex plane with radius $\rho > 0$. Then $S'_{n+1,m+1}(\theta)$ and $T'_{n+1,m+1}(\theta)$ have representations as contour integrals*

$$
\begin{aligned}
S'_{n+1,m+1}(\theta) &= \frac{\theta^m}{(\theta+1)_n} \frac{1}{2\pi i} \int_{\mathcal{C}_\rho} \frac{(z+1)_n}{z^m} \frac{dz}{z-\theta}, \quad \rho > \theta, \\
T'_{n+1,m+1}(\theta) &= \frac{\theta^m}{(\theta+1)_n} \frac{1}{2\pi i} \int_{\mathcal{C}_\rho} \frac{(z+1)_n}{z^m} \frac{dz}{\theta-z}, \quad \rho < \theta.
\end{aligned}
\tag{3.2}
$$

Here, n and m are positive integers, $0 \le m \le n$, and θ is a real positive number. The symbol $(\alpha)_n$ denotes the Pochhammer symbol introduced in (1.3).

The main asymptotic results follow from representations given in [2] and are summarised in the next theorem.

Theorem 3.2. *$S'_{n+1,m+1}(\theta)$ and $T'_{n+1,m+1}(\theta)$ have the representations*

$$
\begin{aligned}
S'_{n+1,m+1}(\theta) &= I_x(m, n-m+1) + R'_{n+1,m+1}(\theta), \qquad x = \frac{\tau}{1+\tau}, \\
T'_{n+1,m+1}(\theta) &= I_{1-x}(n-m+1, m) - R'_{n+1,m+1}(\theta), \qquad 1-x = \frac{1}{1+\tau},
\end{aligned}
\tag{3.3}
$$

where $I_x(p, q)$ is the incomplete beta function defined by

$$
I_x(p, q) = \frac{1}{B(p, q)} \int_0^x t^{p-1}(1-t)^{q-1}\, dt,
\tag{3.4}
$$

with

$$
0 < x < 1, \quad p > 0, \quad q > 0, \quad B(p, q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}.
\tag{3.5}
$$

The term $R'_{n+1,m+1}(\theta)$ can be expanded in negative powers of the large parameter n, as will be explained in later sections. For the relation between τ and θ we refer to Definition 3.4.

Observe that in the representations given in this theorem the complementary relation in (2.6) is preserved because of the complementary property of the incomplete beta function:

$$
I_x(p, q) = 1 - I_{1-x}(q, p).
\tag{3.6}
$$

3.1   The asymptotic approach

The representations given in Theorem 3.2 are obtained in [2] by using the saddle point method. The first step is the introduction of a phase function $\phi(z)$. We write

$$
S'_{n+1,m+1}(\theta) = \frac{e^{-\phi(\theta)}}{2\pi i} \int_{\mathcal{C}_\rho} e^{\phi(z)} \frac{dz}{z-\theta}, \quad \rho > \theta,
\tag{3.7}
$$

where $\mathcal{C}_\rho$ is a circle as in Theorem 3.1 and

$$
\begin{aligned}
\phi(z) &= \ln\left((z+1)_n\right) - m \ln z = \sum_{k=0}^{n-1} \ln(z+1+k) - m \ln z \\
&= \ln\Gamma(z+1+n) - \ln\Gamma(z+1) - m \ln z.
\end{aligned}
\tag{3.8}
$$

For positive values of z, we have the limiting forms

$$\phi(z) \sim -m \ln z, \quad z \to 0; \qquad \phi(z) \sim n \ln(z+1), \quad z \to \infty, \qquad (3.9)$$

where the second estimate comes from $\Gamma(z+1+n)/\Gamma(z+1) \sim (z+1)^n$. In addition, there is one positive minimum z_0 of $\phi(z)$. A proof is given in [8].

The function

$$\chi(t) = n \ln(t+1) - m \ln t, \quad t > 0, \qquad (3.10)$$

has the same limiting behaviour as $\phi(z)$ at $t = 0$ and as $t \to \infty$, and it has one positive minimum $t_0 = m/(n-m)$. In fact, these functions behave quite similar for positive values of their arguments, and we have the following lemma.

Lemma 3.3. *Consider for positive z and t the equation*

$$\phi(z) - \phi(z_0) = \chi(t) - \chi(t_0). \qquad (3.11)$$

Then there is a one-to-one relation between z and t when we use the following condition: $\operatorname{sign}(z - z_0) = \operatorname{sign}(t - t_0)$.

Proof. In Figure 1 of [2] we have drawn graphs of both functions for $m = 38$ and $n = 100$. Both derivaties of the non-negative convex functions have a unique positive zero z_0 and t_0 and their convex curves touch the positive real axis at z_0 and t_0. The sign condition for the relation in (3.11) means that the left branches of the curves correspond with functions values for $z \in (0, z_0]$ and $t \in (0, t_0]$, and the right branches with values for $z \in [z_0, \infty)$ and $t \in [t_0, \infty)$. Clearly, we can uniquely determine $z(t)$ and $t(z)$ for positive values of these parameters. $\qquad \square$

Before we start with deriving the asymptotic representations given in Theorem 3.2 we define the following special points used in this paper.

Definition 3.4. Special points

1. *The point z_0, the positive minimum of $\phi(z)$ and the positive solution of the equation $\phi'(z) = 0$, where*

$$\phi'(z) = \sum_{k=0}^{n-1} \frac{1}{z+1+k} - \frac{m}{z}$$
$$= \psi(z+n+1) - \psi(z+1) - \frac{m}{z}, \quad \psi(z) = \frac{\Gamma'(z)}{\Gamma(z)}, \qquad (3.12)$$

   is called the saddle point *of the integral in (3.7).*

2. *The value θ_t for which $S'_{n,m}(\theta_t) = T'_{n,m}(\theta_t) = \frac{1}{2}$ is called the* transition point *of $S'_{n,m}(\theta)$ and $T'_{n,m}(\theta)$.*

3. *The positive value of t that satisfies the relation in (3.11) when z is replaced by θ is called τ. That is, when we write the general solution of (3.11) as $t(z)$, then $\tau = t(\theta)$. Also, τ is the positive solution of*

$$\phi(\theta) - \phi(z_0) = \chi(\tau) - \chi(t_0), \quad \operatorname{sign}(\theta - z_0) = \operatorname{sign}(\tau - t_0). \qquad (3.13)$$

The relation in (3.11) will be used as a transformation of variables (which is also used in [8]) and (3.7) becomes

$$S'_{n+1,m+1}(\theta) = \frac{e^{-\chi(\tau)}}{2\pi i} \int_{\mathcal{C}_\sigma} \frac{(t+1)^n}{t^m} f(t)\,dt,$$

$$f(t) = \frac{1}{z-\theta}\frac{dz}{dt}, \quad \frac{dz}{dt} = \frac{\chi'(t)}{\phi'(z)}, \quad \chi'(t) = (n-m)\frac{t-t_0}{t(1+t)}, \tag{3.14}$$

with $t_0 = m/(n-m)$, where we have used the relation for $\chi(\tau)$ in (3.13). The value $z = \theta$ (a pole of the integrand in (3.7)) corresponds with $t = \tau$ (see Definition 3.4), and this means that the function $f(t)$ will have a pole at $t = \tau$.

The main asymptotic result of [2] is given in the following theorem.

Theorem 3.5. *Let the function $g(t)$ be defined by*

$$g(t) = f(t) - \frac{1}{t-\tau}, \tag{3.15}$$

where $f(t)$ is defined in (3.14). Then the function $R'_{n+1,m+1}(\theta)$ of the representations given in (3.3), has the integral representation

$$R'_{n+1,m+1}(\theta) = \frac{e^{-\chi(\tau)}}{2\pi i} \int_{\mathcal{C}_\sigma} \frac{(t+1)^n}{t^m} g(t)\,dt, \tag{3.16}$$

where $\mathcal{C}_\sigma$ is a contour around the origin and inside the domain where $g(t)$ is analytic. A first-order approximation is given by

$$R'_{n+1,m+1}(\theta) \sim e^{-\chi(\tau)}\binom{n}{m-1} g(t_0), \quad t_0 = \frac{m}{n-m}, \tag{3.17}$$

where

$$g(t_0) = f(t_0) - \frac{1}{t_0-\tau}, \quad f(t_0) = \frac{1}{z_0-\theta}\sqrt{\frac{\chi^{(2)}(t_0)}{\phi^{(2)}(z_0)}}. \tag{3.18}$$

The incomplete beta function (see (3.4)) is used with the representation

$$I_{\frac{\tau}{1+\tau}}(m, n-m+1) = \frac{e^{-\chi(\tau)}}{2\pi i} \int_{\mathcal{C}_\sigma} \frac{(t+1)^n}{t^m} \frac{dt}{t-\tau}. \tag{3.19}$$

This function has the representation (see [6, §8.17(i)])

$$I_{\frac{\tau}{1+\tau}}(m, n-m+1) = (1+\tau)^{-n} \sum_{j=m}^{n} \binom{n}{j} \tau^j, \tag{3.20}$$

and from the complementary relation in (3.6) it follows that $I_{\frac{1}{1+\tau}}(n-m+1, m)$ used in (3.3) has the expansion

$$I_{\frac{1}{1+\tau}}(n-m+1, m) = (1+\tau)^{-n} \sum_{j=0}^{m-1} \binom{n}{j} \tau^j. \tag{3.21}$$

Remark 3.6. The role of the transition point θ_t in connection with Fu's F_s is explained before (2.8). The transition value θ_t can be obtained by using the inversion methods of §4. On the other hand, the main term in the first line of (3.3) is the incomplete beta function, which function has [5] the transition point x_t close to $x = p/(p+q)$, which gives $\tau = m/(n+m+1)$. This point is close to $t_0 = m/(n+m)$, the zero of $\chi'(t)$, see (3.14) and saddle point of the integral in (3.14), which corresponds to z_0, the saddle point of integral in (3.7). We conclude that the saddle point z_0 is a good approximation of the transition value θ_t.

4 The inversion problem

We consider the following inversion problem: let m and n be given, together with a value $s \in (0,1)$. Then find θ such that

$$S'_{n,m}(\theta) = s. \tag{4.1}$$

Since $S'_{n,m}(0) = 0$ and $S'_{n,m}(\theta) \to 1$ as $\theta \to \infty$, and $S'_{n,m}(\theta)$ is an increasing function of θ (see also Figure 1 (**Right**)), there is a unique solution θ of this problem.

We consider two approaches to solve this problem, in the first one we use Newton iteration and the other one is especially useful when the parameters n and m are large enough to use asymptotic approximations.

In both methods we use a starting value θ_0 that follows from the value x that solves the reduced equation

$$I_x(p,q) = s, \quad p = m, \quad q = n - m + 1, \quad x = \frac{\tau}{1+\tau}, \tag{4.2}$$

where $I_x(p,q)$ is the incomplete beta function used in (3.3). With this value x we compute $\tau = x/(1-x)$ and the initial value θ_0 in the Newton method then follows from the relation between τ and θ as explained in Definition 3.4. We use the reduced equation because we consider the incomplete beta function as the main asymptotic approximant of $S'_{n+1,m+1}(\theta)$. Also for small values of n and m it gives a useful initial value θ_0.

The inversion of the incomplete beta function is extensively considered in the literature. For an approach for large variable p and q, see [7] while in [5] a fourth order fixed point method and several other approaches are discussed. For an overview of the inversion of other classical cumulative distribution functions, we refer to [9, Chapter 42].

Remark 4.1. The inversion $S'_{n,m}(\theta) = s$ can be replaced by the equation for the complementary function: $T'_{n,m}(\theta) = 1 - s$, which is relevant when $s \sim 1$, and even more relevant when $s = 1 - \sigma$, when σ is known in detail as a small positive number.

Remark 4.2. The inversion of Fu's F_s (see(1.4)), that is, solving the equation $F_s = f$, $f \in \mathbb{R}$, follows immediately from our methods for solving $S'_{n,m}(\theta) = s$. The equation $F_s = f$ is equivalent with solving

$$S'_{n,m}(\theta) = \frac{e^f}{1+e^f}, \quad T'_{n,m}(\theta) = \frac{1}{1+e^f}. \tag{4.3}$$

When f is a large positive number, it is very relevant to solve the second equation.

4.1 The iterative inversion method

When solving the equation in (4.1) with the Newton iterative method we compute a sequence of vales θ_j, $j = 0, 1, \ldots$, from the scheme

$$\theta_{j+1} = \theta_j - \frac{f(\theta_j) - s}{f'(\theta_j)}, \quad j = 0, 1, 2, \ldots, \quad f(\theta) = S'_{n,m}(\theta). \tag{4.4}$$

The starting value θ_0 is obtained from the reduced equation in (4.2). The derivative of $S'_{n,m}(\theta)$ follows from the following lemma.

Theorem 4.3.

$$\frac{d}{d\theta} S'_{n,m}(\theta) = -S'_{n,m}(\theta) \sum_{k=0}^{n-1} \frac{1}{k+\theta} + \widehat{S}'_{n,m}(\theta), \tag{4.5}$$

where

$$\widehat{S}'_{n,m}(\theta) = \frac{1}{(\theta)_n} \sum_{k=m}^{n} (-1)^{n-k} k S_n^{(k)} \theta^{k-1}. \tag{4.6}$$

These functions satisfy the recurrence relation

$$(\theta + n)\widehat{S}'_{n+1,m}(\theta) = n\widehat{S}'_{n,m}(\theta) + \theta\widehat{S}'_{n,m-1}(\theta) + S'_{n,m-1}(\theta), \quad 1 \le m \le n. \tag{4.7}$$

Proof. First we have

$$\frac{d}{d\theta} \frac{1}{(\theta)_n} = \frac{d}{d\theta} \frac{\Gamma(\theta)}{\Gamma(\theta + n)} = \frac{\Gamma'(\theta)}{\Gamma(\theta + n)} - \frac{\Gamma(\theta)\Gamma'(\theta + n)}{\Gamma^2(\theta + n)}. \tag{4.8}$$

Next we use the ψ-function, defied by $\psi(z) = \Gamma'(z)/\Gamma(z)$, which has the recursive property (which is also used in (3.12))

$$\psi(z + n) = \psi(z) + \sum_{k=0}^{n-1} \frac{1}{k+z}, \quad n = 1, 2, 3, \ldots. \tag{4.9}$$

This recursion easily follows from the fundamental property of the gamma function $\Gamma(z + 1) = z\Gamma(z)$. We find

$$\frac{d}{d\theta} \frac{1}{(\theta)_n} = (\psi(\theta) - \psi(\theta + n)) \frac{1}{(\theta)_n}. \tag{4.10}$$

Combining these results we find the relation in (4.5). The proof of the recurrence relation in (4.7) follows from the recurrence relation of the Stirling numbers in (2.3), just as in the proof of Theorem 2.1. $\qquad\square$

A few first values of $\widehat{S}'_{n,m}(\theta)$ are

$$\widehat{S}'_{0,0}(\theta) = 0, \quad \widehat{S}'_{1,0} = \frac{1}{\theta}, \quad \widehat{S}'_{1,1} = \frac{1}{\theta},$$
$$\widehat{S}'_{2,0}(\theta) = \frac{2\theta + 1}{\theta(\theta + 1)}, \quad \widehat{S}'_{2,1} = \frac{2\theta + 1}{\theta(\theta + 1)}, \quad \widehat{S}'_{2,2} = \frac{2}{\theta + 1}. \tag{4.11}$$

In §4.3 we give numerical examples of the Newton iteration scheme.

4.2 The asymptotic inversion method

We consider the inversion of the full equation (4.1) with the representation of $S'_{n+1,m+1}(\theta)$ as given in (3.3). We concentrate on finding τ and with this information we compute θ; see Definition 3.4. We propose the following

Proposition 4.4. *Let x be the solution of the reduced equation in (4.2), with corresponding τ value $\tau_0 = x/(1-x)$. Then we will construct an expansion of the wanted value τ of the form*

$$\tau = \tau_0 + \varepsilon, \quad \varepsilon \sim \frac{\tau_1}{\nu} + \frac{\tau_2}{\nu^2} + \dots, \quad \nu = n - m, \tag{4.12}$$

with

$$\tau_1 = \frac{\tau_0(\tau_0 + 1)}{\tau_0 - t_0} \ln\big((t_0 - \tau_0)f(t_0)\big), \tag{4.13}$$

where $f(t_0)$ is given in (3.18), and

$$e^{-\xi}\tau_2 = \tau_1 \frac{e^{-\xi} - 1}{\xi} + \frac{(2\tau_0 + 1)\tau_1^2}{\tau_0(\tau_0 + 1)} \frac{e^{-\xi} - 1 + \xi}{\xi^2} +$$
$$\rho'(\tau_0)\tau_1^3 \frac{e^{-\xi} - 1 + \xi - \frac{1}{2}\xi^2}{\xi^3} - \tag{4.14}$$
$$\tau_0(\tau_0 + 1)\big(G_1(t_0) + \tau_1 G_0'(t_0) + \tfrac{1}{2}\rho'(\tau_0)\tau_1^2 G_0(t_0)\big).$$

The coefficients $G_k(t)$ are defined in Theorem 5.1 (see also §5) and

$$\rho(\tau) = \frac{t_0 - \tau}{\tau(1 + \tau)}, \quad \xi = \tau_1\rho(\tau_0). \tag{4.15}$$

To obtain these coefficients τ_j we have used a perturbation method that starts with writing $S'_{n+1,m+1}(\theta)$ of Theorem 3.2 in the form

$$I_{\frac{\tau_0 + \varepsilon}{1 + \tau_0 + \varepsilon}}(p, q) + e^{-\chi(\tau_0 + \varepsilon)}\binom{n}{m-1}S(\tau_0 + \varepsilon) = s, \tag{4.16}$$

where $S(\tau)$ is the function with expansion (see (5.1) and (3.3))

$$S(\tau) \sim \sum_{k=0}^{\infty} \frac{G_k(t_0)}{\nu^k}. \tag{4.17}$$

The idea is to use the expansion of ε given in (4.12) in (4.16), and expand the relevant terms in negative powers of ν. The coefficients of the same powers of ν in this collection should vanish. This yields equations for the coefficients τ_j. Because we already calculated τ_0 such that $I_{\frac{\tau_0}{1+\tau_0}}(p, q) = s$, we have the asymptotic equality

$$\sum_{k=1}^{\infty} \frac{\varepsilon^k}{k!} \frac{d^k}{d\tau^k} I_{\frac{\tau}{1+\tau}}(p, q) + \binom{n}{m-1} \sum_{k=0}^{\infty} \frac{\varepsilon^k}{k!} \frac{d^k}{d\tau^k}\left(e^{-\chi(\tau)}S(\tau)\right) = 0, \tag{4.18}$$

where the derivatives are evaluated at $\tau = \tau_0$. The construction of the coefficients τ_j can be done with the help of symbolic calculations. The technical details of the manipulations to find (4.13) and (4.14) are available from the authors.

Table 1: Results for inverting equation $S'_{n,m}(\theta) = s$ by using Newton iteration with starting value $\theta = \theta_0$ and relative errors $\delta_j = |s/S'_{n,m}(\theta_j) - 1|$, $j = 0, 2, 4$.

m/n	s	θ_0	δ_0	θ_2	δ_2	θ_4	δ_4
10/25	0.0001	0.02	0.82	0.812	0.20e-00	0.78467	0.17e-03
10/25	0.25	4.55	0.36	3.786	0.46e-07	3.78618	0.00e-00
10/25	0.50	6.13	0.23	5.163	0.46e-03	5.16527	0.10e-14
10/25	0.75	8.21	0.12	6.970	0.26e-02	6.98945	0.98e-10
25/50	0.0001	6.20	0.64	5.70	0.46e-01	5.67813	0.40e-06
25/50	0.25	16.06	0.24	14.941	0.91e-05	14.9416	0.10e-14
25/50	0.50	19.70	0.15	18.373	0.18e-04	18.3727	0.14e-14
25/50	0.75	24.14	0.080	22.563	0.19e-03	22.5663	0.10e-14

Example 4.5. We take $n = 99$, $m = 49$ and try to find the value θ such that $S'_{100,50}(\theta) = \frac{1}{2}$. This means, we try to find the transition value θ_t for these m and n; see Definition 3.4. This example corresponds with the second line in Table 2. For the asymptotic method we have the following steps.

1. Compute the saddle point $z_0 \doteq 39.1327$ by solving the equation $\phi'(z) = 0$, see (3.12), and $t_0 = 49/50 = 0.98$ from $\chi'(t) = 0$, see (3.14).

2. Compute $\phi(z_0) \doteq 259.198$ and $\chi(t_0) \doteq 68.6165$.

3. With $s = \frac{1}{2}$ solve the equation $I_x(p, q) = \frac{1}{2}$, see (4.2). We find $x \doteq 0.4899330675$. This gives $\tau_0 = x/(1-x) \doteq 0.960527$ and $\chi(\tau_0) \doteq 68.6215$.

4. Use (3.13) to compute θ from the equation $\phi(\theta) = \phi(z_0) + \chi(\tau_0) - \chi(t_0) \doteq 259.203$ in the interval $0, z_0$) (because $\tau_0 < t_0$, and find $\theta \doteq 38.29722$.

5. A first check: compute $S'_{100,50}(\theta)$ with this value of θ and find 0.50233, with relative error 0.0047.

6. Next compute τ_1 from (4.13), with the just found value of θ that is needed in $f_0 = f(t_0)$ given in (3.18). Find $\tau_1 \doteq -0.055873923$ and compute $\tau \sim \tau_0 + \tau_1/\nu \doteq 0.959409535$, with $\nu = n - m$.

7. Repeat the steps given above: the equation for the new θ becomes $\phi(\theta) = \phi(z_0) + \chi(\tau_0 + \tau_1/\nu) - \chi(t_0) \doteq 259.203352$ and find $\theta \doteq 38.2492993$.

8. Check: compute $S'_{100,50}(\theta)$ with this θ and find 0.5000190, with relative error 0.38e-4.

9. With the next term in the expansion, $\tau \sim \tau_0 + \tau_1/\nu + \tau_2/\nu^2$, we find $\theta \doteq 38.248908191$, and $S'_{100,50}(\theta) \doteq 0.500000125$, with relative error 0.25e-6.

Table 2: Results for computing θ from the equation $S'_{nm}(\theta) = s$ by using the approximation of τ in (4.12) with terms up to τ_j, $j = 0, 1, 2$; see (4.13) and (4.14). We give the corresponding values θ_j and relative errors $\delta_j = |s/S'_{nm}(\theta) - 1|$.

m/n	s	θ_0	δ_0	θ_1	δ_1	θ_2	δ_2
200/250	0.0001	255.3	0.36e-2	255.339	0.24e-4	255.33835	0.13e-4
200/250	0.25	408.2	0.11e-2	408.103	0.35e-5	408.10264	0.66e-7
200/250	0.50	455.0	0.66e-3	454.911	0.21e-5	454.91098	0.18e-6
200/250	0.75	508.2	0.34e-3	508.124	0.11e-5	508.12328	0.76e-8
500/1000	0.0001	307.4	0.73e-2	307.383	0.58e-5	307.38266	0.32e-8
500/1000	0.25	378.6	0.23e-2	378.570	0.19e-5	378.56980	0.10e-8
500/1000	0.50	396.4	0.15e-2	396.387	0.11e-5	396.39298	0.20e-3
500/1000	0.75	415.1	0.77e-3	415.025	0.62e-6	415.02539	0.20e-9

4.3 Numerical results for the inversion

In Table 1 we give the results for computing θ from the equation $S'_{n,m}(\theta) = s$ by using Newton iteration (see §4.1). The starting value θ_0 is obtained by inverting the reduced equation in (4.2), where also the relation between x and θ is explained. In the table we give the iterated values of θ_j and $\delta_j = |s/S'_{n,m}(\theta_j) - 1|$ for $j = 0, 2, 4$. As can be expected by using Newton iteration, once we have a reasonable starting value, we can obtain excellent accuracy with a few iteration steps. We even see convergence for small values $s = 0.0001$, where for the corresponding θ values the curve of $S'_{n,m}(\theta)$ is very flat (see Figure 1, **Right**).

In Table 2 we give the results for computing θ by using asymptotic methods described in §4.2. The equation $S'_{n+1,m+1}(\theta) = s$ is approximately solved by using the approximation of τ in (4.12) with terms up to τ_j, $j = 0, 1, 2$. As expected, we see a better performance when we include τ_1/ν and $\tau_1/\nu + \tau_2/\nu^2$.

5 Deriving the complete asymptotic expansion

The first-term approximation in our previous result in (3.16) of Theorem 3.5 will now be extended in the following theorem.

Theorem 5.1. *Let $R'_{n+1,m+1}(\theta)$ be defined as in (3.16). Then for $N = 0, 1, 2, \ldots$*

$$R'_{n+1,m+1}(\theta) = e^{-\chi(\tau)}\binom{n}{m-1}\sum_{k=0}^{N-1}\frac{G_k(t_0)}{\nu^k} + \frac{e^{-\chi(\tau)}}{2\pi i\nu^N}\int_{\mathcal{C}_\sigma}\frac{(t+1)^n}{t^m}G_N(t)\,dt, \quad (5.1)$$

where $\nu = n - m$, $G_0(t) = g(t)$, see (3.15), and other $G_k(t)$ follow from the recursive

scheme

$$H_k(t) = \frac{G_k(t) - G_k(t_0)}{t - t_0}, \quad G_{k+1}(t) = -\frac{d}{dt}\Big(t(1+t)H_k(t)\Big), \quad k = 0, 1, 2, \dots . \tag{5.2}$$

Proof. We start with the representation in (3.16) and use an integration by parts procedure, which starts by writing

$$G_0(t) = G_0(t_0) + (t - t_0)H_0(t), \quad G_0(t) = g(t). \tag{5.3}$$

This gives

$$R'_{n+1,m+1}(\theta) = G_0(t_0)e^{-\chi(\tau)}\binom{n}{m-1} + \frac{e^{-\chi(\tau)}}{2\pi i}\int_{\mathcal{C}_\sigma} \frac{(t - t_0)H_0(t)}{\chi'(t)}\, de^{\chi(t)}, \tag{5.4}$$

and using $\chi'(t)$ shown in (3.14), we find

$$R'_{n+1,m+1}(\theta) = G_0(t_0)e^{-\chi(\tau)}\binom{n}{m-1} + \frac{e^{-\chi(\tau)}}{2\pi i\nu}\int_{\mathcal{C}_\sigma} \frac{(t + 1)^n}{t^m}G_1(t)\, dt, \tag{5.5}$$

where

$$G_1(t) = -\frac{d}{dt}\Big(t(1 + t)H_0(t)\Big). \tag{5.6}$$

This integral has the same form as the one in (3.16), and we can continue this method. This proves the theorem.

$\square$

The first coefficients $G_k(t_0)$ of the expansion in (5.1) are

$$G_0(t_0) = g_0, \quad G_1(t_0) = -(1 + 2t_0)g_1 - t_0(t_0 + 1)g_2,$$

$$G_2(t_0) = 2(1 + 2t_0)g_1 + (2 + 11t_0 + 11t_0^2)g_2 + \tag{5.7}$$

$$5t_0(t_0 + 1)(1 + 2t_0)g_3 + 3t_0^2(t_0 + 1)^2 g_4.$$

The coefficients g_k follow from the coefficients f_k in the expansion $f(t) = \sum_{k=0}^{\infty} f_k(t - t_0)^k$, with $f(t)$ defined in (3.14). We have

$$g(t) = f(t) - \frac{1}{t - \tau} = \sum_{k=0}^{\infty} g_k(t - t_0)^k \quad g_k = f_k - \frac{(-1)^k}{(t_0 - \tau)^{k+1}}. \tag{5.8}$$

Finally, all these coefficients can be expressed in terms of the coefficients z_k of the expansion

$$z - z_0 = \sum_{k=1}^{\infty} z_k(t - t_0)^k, \tag{5.9}$$

13

and these follow from substituting this expansion in the Taylor expansions of the functions used in the transformation given in (3.11). This transformation has the local expansions at the saddle points

$$(z - z_0)\sqrt{\sum_{k=2}^{\infty} \frac{1}{k!}\phi^{(k)}(z_0)(z - z_0)^{k-2}} = (t - t_0)\sqrt{\sum_{k=2}^{\infty} \frac{1}{k!}\chi^{(k)}(t_0)(t - t_0)^{k-2}}, \quad (5.10)$$

where the square roots are positive for positive values of z and t. The derivatives $\phi^{(k)}(z)$ can be expressed in terms of the derivatives of the gamma functions; see (3.12). The derivatives $\chi^{(k)}(t)$ at $t = t_0$ are simple expressions.

The first coefficients z_k are

$$z_1 = \sqrt{\frac{\chi^{(2)}(t_0)}{\phi^{(2)}(z_0)}}, \quad z_2 = \frac{\chi^{(3)}(t_0) - z_1^3 \phi^{(3)}(z_0)}{6z_1 \phi^{(2)}(z_0)}. \quad (5.11)$$

The first coefficients f_k are

$$f_0 = \frac{z_1}{z_0 - \theta}, \quad f_1 = \frac{2z_2 z_0 - 2z_2\theta - z_1^2}{(\theta - z_0)^2}$$

$$f_2 = \frac{6z_3 z_0 \theta + 3z_1 z_2 z_0 - 3z_1 z_2 \theta - z_1^3 - 3z_3 z_0^2 - 3z_3 \theta^2}{(\theta - z_0)^3}. \quad (5.12)$$

Then, the first coefficients g_k follow from (5.8).

5.1 Numerical verifications of the asymptotic approximation

A convenient tool for verifying the errors in numerical calculations is the recursion in Theorem 2.1, (2.4). We can write this in the form

$$\frac{nS'_{n,m}(\theta) + \theta S'_{n,m-1}(\theta)}{(\theta + n)S'_{n+1,m}(\theta)} - 1 = 0. \quad (5.13)$$

Especially for large values of n (for example, $n = 100.000$) used in the tests we avoid the exact evaluation of the sums in (1.1), and we accept that we do not verify a standard relative error.

In Table 3 we show the values of the relation (5.13) in the computation of $S'_{n,m}(\theta)$ for $n = 1000$ and $n = 100.000$ for several values of m. In the first two parts ($n = 1.000$) of the table we used the expansion in (5.1) with terms up to and including $k = 3$, and in the final two parts ($n = 100.000$) we only used the terms with $G_0(t_0)$ and $G_1(t_0)$. We have taken $\theta = \rho z_0$, for the shown values of ρ. In this way the values of $S'_{n,m}(\theta)$ are not very small. For example, with $n = 1000.000$, $m = 75.000$, $z_0 \doteq 136312.21$, we have

$$\rho = 0.97 \quad \Longrightarrow \quad S'_{n,m}(\theta) \doteq 0.300778124649e\text{-}04,$$

$$\rho = 1.00 \quad \Longrightarrow \quad S'_{n,m}(\theta) \doteq 0.501722781430e\text{-}00. \quad (5.14)$$

Table 3: Values of the relation (5.13) in the computation of $S'_{n,m}(\theta)$ for $n = 1.000$ and $n = 100.000$ for several values of m and θ.

$n = 1.000$	Digits $= 16$	4 terms				
$\rho \setminus m$	150	300	450	600	750	900
0.70	$0.16e$-12	$0.27e$-13	$0.73e$-13	$0.10e$-12	$0.12e$-12	$0.47e$-13
0.80	$0.32e$-12	$0.13e$-12	$0.16e$-12	$0.86e$-13	$0.11e$-12	$0.88e$-13
0.90	$0.24e$-11	$0.30e$-12	$0.70e$-13	$0.60e$-13	$0.50e$-12	$0.13e$-10
1.00	$0.12e$-11	$0.19e$-11	$0.14e$-12	$0.54e$-12	$0.29e$-11	$0.79e$-13
$n = 1.000$	Digits $= 20$	4 terms				
$\rho \setminus m$	150	300	450	600	750	900
0.70	$0.78e$-15	$0.58e$-16	$0.17e$-17	$0.49e$-16	$0.16e$-16	$0.24e$-16
0.80	$0.47e$-15	$0.48e$-16	$0.93e$-17	$0.16e$-16	$0.60e$-18	$0.20e$-16
0.90	$0.12e$-15	$0.36e$-16	$0.28e$-16	$0.60e$-17	$0.56e$-16	$0.20e$-14
1.00	$0.17e$-15	$0.14e$-15	$0.93e$-16	$0.37e$-16	$0.37e$-16	$0.95e$-16
$n = 100.000$	Digits $= 16$	2 terms				
$\rho \setminus m$	15.000	30.000	45.000	60.000	75.000	90.000
0.97	$0.53e$-10	$0.36e$-10	$0.86e$-11	$0.82e$-11	$0.24e$-11	$0.47e$-11
0.98	$0.58e$-11	$0.12e$-10	$0.61e$-11	$0.14e$-10	$0.16e$-11	$0.17e$-10
0.99	$0.86e$-11	$0.17e$-10	$0.18e$-10	$0.16e$-10	$0.20e$-10	$0.44e$-11
1.00	$0.34e$-09	$0.28e$-08	$0.47e$-09	$0.14e$-08	$0.46e$-11	$0.19e$-08
$n = 100.000$	Digits $= 20$	2 terms				
$\rho \setminus m$	15.000	30.000	45.000	60.000	75.000	90.000
0.97	$0.20e$-14	$0.21e$-14	$0.16e$-14	$0.16e$-14	$0.63e$-15	$0.97e$-15
0.98	$0.37e$-14	$0.38e$-14	$0.15e$-14	$0.29e$-16	$0.80e$-15	$0.25e$-15
0.99	$0.85e$-14	$0.10e$-14	$0.18e$-14	$0.14e$-14	$0.16e$-14	$0.64e$-15
1.00	$0.71e$-13	$0.22e$-12	$0.23e$-13	$0.15e$-13	$0.18e$-12	$0.51e$-13

The computations are done with Maple 2020, Digits=16. To show the effect of cancellation or rounding errors, we have used the same Maple codes with Digits=20.

The first three coefficients $G_k(t_0)$ of the expansion in (5.1) are given in (5.7). Each $G_k(t_0)$ is a linear combination of coefficients g_k, $k = 1, 2, \ldots, 2k$, and these are defined in (5.8). When $\theta \sim \theta_t$, the transition value, $|t_0 - \tau|$ is small, and in the limit $\tau \to t_0$ the coefficient g_k is well defined, although f_k has a pole at $t = t_0$ with corresponding z-value $\theta = z_0$ (see the definition of $f(t)$ in (3.14)). In g_k these poles are cancelled. From an analytical point of view, everything runs fine, but the algorithm needs special attention. For this we use expansions of g_k for small values of $|t_0 - \tau|$ in the form

$$g_k = \sum_{j=0}^{\infty} g_{j,k}(\tau - t_0)^j, \quad k = 0, 1, 2, \ldots . \tag{5.15}$$

The largest errors in the first part and third part (both with Digits=16) in Table 3 occur for $\rho = 1$, that is, when $\theta = z_0$, hence $\theta \sim \theta_t$, the transition value. In this case $\tau \sim t_0$. This is related to the difficulty of computing the coefficients $G_k(t_0)$ near the transition point θ_t, which is near z_0, as explained in Definition 3.4 ad Remark 3.6. This is an interesting analytical issue, but for the population genetics application area it is less important. In fact we accept some loss of accuracy near the transition point instead of using more complicated analytical expansions of the coefficients.

6   Conclusions

We have provided additional details for the asymptotic approximation of the cumulative distribution quantities $S'_{n,m}(\theta)$ and $T'_{n,m}(\theta) = 1 - S'_{n,m}(\theta)$. As a completely new contribution, we have considered the inversion problem to compute θ, the solution of the equation $S'_{n,m}(\theta) = s$, with given $s \in (0, 1)$, m and large n. A simple inversion method uses Newton iteration, the other one asymptotic approximations. For this we provided additional coefficients in the earlier given asymptotic expansion. We have shown that some loss of accuracy is localized near the transition values. We further observe that these estimation errors can be mitigated with additional terms of the expansion. These errors are largely in a regime where the distribution functions are near the changeover value $\frac{1}{2}$, which is an interesting domain from an analytical point of view, but where Fu's F_s values are not important for genetic inferences.

Acknowledgments

References

[1] S. L. Chen. Implementation of a Stirling number estimator enables direct calculation of population genetics tests for large sequence datasets. *Bioinformatics*, 35(15):2668–2670, 2019.

[2] S. L. Chen and N. M. Temme. A faster and more accurate algorithm for calculating population genetics statistics requiring sums of Stirling numbers of the first kind. *G3: Genes, Genomes, Genetics*, 2020. https://doi.org/10.1534/g3.120.401575.

[3] Y. X. Fu. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics*, 147(2):915–925, Oct 1997.

[4] A. Gil, J. Segura, and N. M. Temme. *Numerical Methods for Special Functions.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2007.

[5] A. Gil, J. Segura, and N. M. Temme. Efficient algorithms for the inversion of the cumulative central beta distribution. *Numer. Algorithms*, 74(1):77–91, 2017.

[6] R. B. Paris. Chapter 8, Incomplete gamma and related functions. In *NIST Handbook of Mathematical Functions*, pages 173–192. U.S. Dept. Commerce, Washington, DC, 2010. http://dlmf.nist.gov/8.

[7] N. M. Temme. Asymptotic inversion of the incomplete beta function. *J. Comput. Appl. Math.*, 41(1-2):145–157, 1992.

[8] N. M. Temme. Asymptotic estimates of Stirling numbers. *Stud. Appl. Math.*, 89(3):233–243, 1993.

[9] N. M. Temme. *Asymptotic methods for integrals*, volume 6 of *Series in Analysis.* World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2015.

[10] Z. Wu, B. Periaswamy, O. Sahin, M. Yaeger, P. Plummer, W. Zhai, Z. Shen, L. Dai, S. L. Chen, and Q. Zhang. Point mutations in the major outer membrane protein drive hypervirulence of a rapidly expanding clone of Campylobacter jejuni. *Proc. Natl. Acad. Sci. U.S.A.*, 113(38):10690–10695, 09 2016.