ISSN 2220-5438

Moscow Journal of Combinatorics and Number Theory

МФТИ

URSS

Volume 1 • Issue 3

2011

Moscow Journal of Combinatorics and Number Theory. 2011. Vol. 1. Iss. 3. 80 p.

The journal was founded in 2010.

Published by the Moscow Institute of Physics and Technology
with the support of Yandex and Microsoft.

The aim of this journal is to publish original, high-quality research articles from a broad range of interests within combinatorics, number theory and allied areas. One volume of four issues is published annually.

Website of our journal
`http://mjcnt.phystech.edu`

E-mail
`mjcnt@phystech.edu`

Address of the Editorial Board

Moscow institute of physics
and technology (state university)
Faculty of Innovations
and High Technology,
Laboratorny Korpus, k. 209,
9, Institutskii pereulok,
Dolgoprudny,
Moscow Region,
Russia,
141700

Адрес редакции

Московский физико-технический
институт (государственный университет)
Факультет инноваций
и высоких технологий
Лабораторный корпус, к. 209,
Институтский переулок, д. 9,
г. Долгопрудный,
Московская область,
Российская Федерация,
141700

URSS Publishers

56, Nakhimovsky Prospekt,
Moscow,
Russia,
117335

Издательство «УРСС»

Нахимовский пр-т, 56
Москва,
Российская Федерация,
117335

Журнал зарегистрирован в Федеральной службе по надзору в сфере массовых коммуникаций, связи и охраны культурного наследия 3 сентября 2010 г. Свидетельство ПИ № ФС77-41900.

Формат 70 × 100/16. Печ. л. 5. Зак. № ПЖ-43.
Отпечатано в ООО «ЛЕНАНД».
117312, Москва, пр-т Шестидесятилетия Октября, 11А, стр. 11.

ISSN 2220-5438

11210 ID 158900

9 785453 000272

Moscow Journal
of Combinatorics
and Number Theory
2011, vol. 1, iss. 3,
pp. 12–24, [pp. 220–232]

Asymptotically normal distribution of some tree families relevant for phylogenetics, and of partitions without singletons

Éva Czabarka (Columbia), Péter L. Erdős (Budapest),
Virginia Johnson (Columbia), Anne Kupczok (Vienna),
László A. Székely (Columbia)

Abstract: P. L. Erdős and L. A. Székely [*Adv. Appl. Math.* **10** (1989), 488–496] gave a bijection between rooted semilabeled trees and set partitions, which specializes to a bijection between phylogenetic trees and set partitions with classes of size $\geqslant 2$. L. H. Harper's results [*Ann. Math. Stat.* **38** (1967), 410–414] on the asymptotic normality of the Stirling numbers of the second kind translate into asymptotic normality of rooted semilabeled trees with given number of vertices, when the number of internal vertices varies. The asymptotic normality of *modified* Stirling numbers of the second kind that enumerate partitions of a fixed set into a given number of classes of size $\geqslant 2$, which is shown in this paper, translates into the asymptotic normality of the number of phylogenetic trees with given number of vertices, when the number of leaves varies. We also obtain the asymptotic normality of the number of phylogenetic trees with given number of leaves and varying number of internal vertices, which is more relevant for phylogeny. By the bijection, this means the asymptotic normality of the number of partitions of $n+m-1$ elements into m classes of size $\geqslant 2$, when n is fixed and m varies. The proofs are adaptations of the techniques of L. H. Harper [ibid.]. We provide asymptotics for the relevant expectations and variances with error term $O(1/n)$.

Keywords: set partition; generating function; tree; phylogeny; asymptotic enumeration; central limit theorem; local limit theorem

AMS Subject classification: 05A15; 05A16; 05A18; 05C05

Received: 13.10.2010; **revised:** 16.06.2011 and 07.09.2011

1. Asymptotic normality

Let $S(a,b)$ $(S^\star(a,b))$ denote the Stirling number of the second kind, i. e. the number of partitions of an a-element set into b classes, each with at least one (two) elements. $S(a,b)$ is the Stirling number of the second kind, and the Bell number (see the sequence A000110 in [17]) is $\sum_b S(a,b) = B_a$. This paper proves central and local limit theorems for the arrays $S^\star(n,k)$ and $T_{n,m} = S^\star(n-1+m,m)$. Such results for the array $S(a,b)$ have been long known [12]. The technique used in this paper is Harper's method [12], who gave a very elegant proof for the asymptotic normality of the array $S(n,k)$. We follow the interpretation of Canfield [2] and Clark [6], who clarified and explained the details of [12] while generalizing his method, although our discussion is somewhat restrictive. These limit theorems are relevant to phylogenetic tree enumeration by the bijection in [7]. We compute the expectations and variances with $O(1/n)$ error term, to support the phylogeneticists who may use our results for approximation.

Let $A(n,j)$ be an array of non-negative real numbers for $j = 1,\dots,d_n$, and define $A_n(x) = \sum_j A(n,j)x^j$. Observe that $\sum_j A(n,j) = A_n(1)$. Let Z_n denote the random variable, for which the probability $\mathcal{P}(Z_n = j) = \dfrac{A(n,j)}{A_n(1)}$. In terms of $A_n(x)$, there is a well-known [6] and easy to verify expression for the expectation and variance of Z_n:

$$\mathcal{E}(Z_n) = \frac{A_n'(1)}{A_n(1)} \quad \text{and} \quad \mathcal{D}^2(Z_n) = \frac{A_n'(1)}{A_n(1)} + \left(\frac{A_n'(x)}{A_n(x)}\right)'\Bigg|_{x=1}. \tag{1}$$

As $\mathcal{E}(Z_n)$ and $\mathcal{D}(Z_n)$ are determined by the array $A(n,j)$, we will also write them as $\mathcal{E}(A(n,.))$ and $\mathcal{D}(A(n,.))$

The array $A(n,j)$ is called *asymptotically normal* in the sense of a *central limit theorem*, if

$$\frac{1}{A_n(1)} \sum_{j=1}^{\lfloor x_n \rfloor} A(n,j) \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt, \tag{2}$$

as $n \to \infty$ uniformly in x, where

$$x_n = \mathcal{E}(Z_n) + x\mathcal{D}(Z_n). \tag{3}$$

If all the roots of the polynomial $A_n(x)$ are non-positive real numbers, and

$$\lim_{n\to\infty} \mathcal{D}(Z_n) = \infty, \tag{4}$$

then the array $A(n, j)$ is known to be asymptotically normal, furthermore, by [2, 13], the following *local limit theorem* holds:

$$\lim_{n\to\infty} \frac{\mathcal{D}(Z_n)}{A_n(1)} A(n, \lfloor x_n \rfloor) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \tag{5}$$

uniformly in x. The fact that the convergence of the $A(n, j)$ numbers to the Gaussian function is actually uniform, implies that the number $k = J_n$ maximizing $A(n, k)$ satisfies

$$J_n - \mathcal{E}(Z_n) = o(\mathcal{D}(Z_n)); \tag{6}$$

and

$$A(n, J_n) \sim \frac{1}{\sqrt{2\pi}} \frac{A_n(1)}{\mathcal{D}_n(Z_n)}. \tag{7}$$

2. Set partitions without singletons

Theorem 1. *For the sequence* $A(n, j) = S^\star(n, j)$ *the central limit theorem* (2) *and the local limit theorem* (5) *hold with*

$$\mathcal{E}(S^\star(n, .)) = \frac{n}{r} - r - \frac{1}{2r} + \frac{1}{2r(r+1)^2} + O\Big(\frac{1}{n}\Big), \tag{8}$$

$$\mathcal{D}^2(S^\star(n, .)) = \frac{n}{r(r+1)} - r + 1 - \frac{2}{r+1} - \frac{1}{2(r+1)^2} - \frac{1}{2(r+1)^3} + \\ + \frac{1}{(r+1)^4} + O\Big(\frac{1}{n}\Big). \tag{9}$$

In explicit terms,

$$\mathcal{E}(S^\star(n, .)) = \frac{n}{\ln n} + \frac{n\big(\ln\ln n + O\big(\frac{1}{\ln n}\big)\big)}{\ln^2 n}, \tag{10}$$

$$\mathcal{D}^2(S^\star(n, .)) = \frac{n}{\ln^2 n} + \frac{n\big(2\ln\ln n - 1 + O\big(\frac{1}{\ln n}\big)\big)}{\ln^3 n}. \tag{11}$$

Furthermore, the number $k = J_n$ *that maximizes* $S^\star(n,k)$ *satisfies*

$$J_n = \frac{n}{r} + o\left(\frac{\sqrt{n}}{r}\right) \tag{12}$$

and

$$S^\star(n, J_n) = \frac{rB_{n-1}}{\sqrt{2n\pi}}(1 + o(1)). \tag{13}$$

Proof. We start with some facts that we need. Set $B_n^\star = \sum_k S^\star(n,k)$, the number of all partitions of an n-element set not using singleton classes (see the sequence A000296 in [17]). Becker [1] observed that [1)]

$$B_n = B_{n+1}^\star + B_n^\star. \tag{14}$$

From $B_i = B_i^\star + B_{i+1}^\star$ for $i = 1, 2, \ldots, n$, and $B_1^\star = 0$, we obtain $B_{n+1}^\star = \sum_{i=1}^{n} B_i(-1)^{n-i}$. As the B_n sequence is strictly increasing, we immediately obtain $B_t - B_{t-1} < B_{t+1}^\star = \sum_{i=1}^{t} B_i(-1)^{t-i} < B_t$ for $t \geqslant 3$, and with $t = n - h$ the asymptotical formula

$$B_{n+1}^\star = B_n - B_{n-1} + \ldots + (-1)^h B_{n-h} + O(B_{n-h-1}). \tag{15}$$

In the special case $h = 0$, using (21), we obtain:

$$B_{n+1}^\star = B_n - O(B_{n-1}) = B_n\left(1 - O\left(\frac{r}{n}\right)\right). \tag{16}$$

We obtain the recurrence relation

$$S^\star(n,k) = (n-1)S^\star(n-2, k-1) + kS^\star(n-1, k), \tag{17}$$

according to the case analysis whether the n^{th} element is in a doubleton class or not. We define the polynomial sequence $S_n(x) = \sum_k S^\star(n,k)x^k$. It is easy to see

[1)] Identity (14) can be proved by the following bijection from the partitions with at least one singleton class of an n-element set, $[n]$, to the partitions without singleton classes of an $n+1$-element set, $[n+1]$: build a new class from the elements of all singletons and $n+1$.

that $S_1(x) = 0$, $S_2(x) = x$, and for $n \geqslant 3$ from (17),

$$S_n(x) = (n-1)xS_{n-2}(x) + xS'_{n-1}(x). \tag{18}$$

For the proof, first we compute $\mathcal{E}(S^\star(n,.))$ and $\mathcal{D}(S^\star(n,.))$ exactly and then asymptotically. The central and local limit theorems hinge on the relation $\mathcal{D}(S^\star(n,.)) \to \infty$. Salvy and Shackell [15] showed (10,11) originally for $S(n,k)$, but not for $S^\star(n,k)$. However, it is shown in [4] that

$$\mathcal{E}(S(n,.)) = \frac{n}{r} - 1 + \frac{r}{2(r+1)^2} + O\left(\frac{1}{n}\right), \tag{19}$$

$$\mathcal{D}^2(S(n,.)) = \frac{n}{r(r+1)} + \frac{r(r-1)}{2(r+1)^4} - 1 + O\left(\frac{r}{n}\right). \tag{20}$$

Using these asymptotic expansions we obtain that $\mathcal{E}(S^\star(n,.)) - \mathcal{E}(S(n,.)) = O(r)$ and $\mathcal{D}^2(S^\star(n,.)) - \mathcal{D}^2(S(n,.)) = O(r)$. The explicit asymptotics follow from these remarkably small differences. Formulae (12) and (13) follow from (6) and (7), where B^*_n is approximated with B_{n-1} by (16).

We obtain from (1), using (18) repeatedly,

$$\begin{aligned}
\mathcal{E}(S^\star(n,.)) &= \frac{B^\star_{n+1}}{B^\star_n} - n\frac{B^\star_{n-1}}{B^\star_n},\\
\mathcal{D}^2(S^\star(n,.)) &= \frac{B^\star_{n+2}}{B^\star_n} + 2n\frac{B^\star_{n+1}B^\star_{n-1}}{(B^\star_n)^2} + n(n-1)\frac{B^\star_{n-2}}{B^\star_n} -\\
&\quad - \left(\frac{B^\star_{n+1}}{B^\star_n}\right)^2 - n^2\left(\frac{B^\star_{n-1}}{B^\star_n}\right)^2 - n\frac{B^\star_{n-1}}{B^\star_n} - (2n+1).
\end{aligned}$$

Canfield and Harper [5], Canfield [3] made minor modifications on the asymptotics of Moser and Wyman [14] to develop asymptotics for B_{n+h}, which holds uniformly for $h = O(\ln n)$, using a *single* $r = r(n)$ value, as $n \to \infty$:

$$\begin{aligned}
B_{n+h} &= \frac{(n+h)!}{r^{n+h}}\frac{e^{e^r-1}}{(2\pi B)^{1/2}} \times \left(1 + \frac{P_0 + hP_1 + h^2P_2}{e^r} + \right.\\
&\quad \left. + \frac{Q_0 + hQ_1 + h^2Q_2 + h^3Q_3 + h^4Q_4}{e^{2r}} + O(e^{-3r})\right),
\end{aligned} \tag{21}$$

where $B = (r^2 + r)e^r$, P_i and Q_i are explicitly known rational functions of r. We list and use in the Maple worksheet [18] their exact values from Canfield [4].

To obtain (8) and (9), we started with the closed forms above, used (15) for the $B^{\star}$ numbers, substituted the B numbers with (21), and changed e^{-r} to r/n. For details, see the Maple worksheet [18].

Finally, Lemma 1 will provide the non-positive real roots of the generating polynomial. By induction from (18) we see that for $n \geqslant 2$ one has

$$\deg(S_n(x)) = \left\lfloor \frac{n}{2} \right\rfloor \tag{22}$$

and the root $x = 0$ has multiplicity one. Hence $S_n'(0) > 0$ for $n \geqslant 2$. □

Lemma 1. *Apart from $x = 0$, the roots of $S_{2n}(x)$ and $S_{2n+1}(x)$ are negative real numbers and every root occurs with multiplicity one. Furthermore, if the roots of $S_{2n}(x)$ are denoted by $\beta_i^{(2n)}$ in increasing order, and the roots of $S_{2n-1}(x)$, $S_{2n+1}(x)$ are denoted by $\alpha_i^{(2n-1)}$, $\alpha_i^{(2n+1)}$, both in increasing order, then the following interlacing properties hold*:

$$\beta_1^{(2n)} < \alpha_1^{(2n-1)} < \beta_2^{(2n)} < \alpha_2^{(2n-1)} < \ldots < \beta_{n-1}^{(2n)} < \alpha_{n-1}^{(2n-1)} = 0 = \beta_n^{(2n)},$$

$$\beta_1^{(2n)} < \alpha_1^{(2n+1)} < \beta_2^{(2n)} < \alpha_2^{(2n+1)} < \ldots < \alpha_{n-2}^{(2n+1)} < \beta_{n-1}^{(2n)} < \alpha_{n-1}^{(2n+1)} < \beta_n^{(2n)} = 0 = \alpha_n^{(2n+1)}.$$

We will use induction in n. The roots of $S_2(x) = S_3(x) = x$, $S_4(x) = 3x^2 + x$ (roots $\beta_1^{(4)} = -1/3$ and $\beta_2^{(4)} = 0$) and $S_5(x) = 10x^2 + x$ (roots $\alpha_1^{(5)} = -1/10$ and $\alpha_2^{(5)} = 0$) satisfy Lemma 1. The inductive step is provided by the following two statements for $n \geqslant 2$:

(i) If the roots of $S_{2n-2}(x)$ and $S_{2n-1}(x)$ occur with multiplicity one and satisfy

$$\beta_1^{(2n-2)} < \alpha_1^{(2n-1)} < \beta_2^{(2n-2)} < \alpha_2^{(2n-1)} < \ldots < \alpha_{n-2}^{(2n-1)} < \beta_{n-1}^{(2n-2)} = 0 = \alpha_{n-1}^{(2n-1)},$$

then the roots $\beta_i^{(2n)}$ of $S_{2n}(x)$ satisfy

$$\beta_1^{(2n)} < \alpha_1^{(2n-1)} < \beta_2^{(2n)} < \alpha_2^{(2n-1)} < \ldots < \beta_{n-1}^{(2n)} < \alpha_{n-1}^{(2n-1)} = 0 = \beta_n^{(2n)}.$$

(ii) If the roots of $S_{2n-1}(x)$ and $S_{2n}(x)$ occur with multiplicity one and satisfy

$$\beta_1^{(2n)} < \alpha_1^{(2n-1)} < \beta_2^{(2n)} < \alpha_2^{(2n-1)} < \ldots < \beta_{n-1}^{(2n)} < \alpha_{n-1}^{(2n-1)} = 0 = \beta_n^{(2n)},$$

then the roots $\alpha_i^{(2n+1)}$ of $S_{2n+1}(x)$ satisfy

$$\beta_1^{(2n)} < \alpha_1^{(2n+1)} < \beta_2^{(2n)} < \alpha_2^{(2n+1)} < \ldots < \alpha_{n-2}^{(2n+1)} < \beta_{n-1}^{(2n)} < \alpha_{n-1}^{(2n+1)} < \beta_n^{(2n)} = 0 = \alpha_n^{(2n+1)}.$$

First we prove (**i**). In our setting the identity (18) specifies to

$$\frac{S_{2n}(x)}{x} = (2n-1)S_{2n-2}(x) + S'_{2n-1}(x), \tag{23}$$

where the RHS is the sum of two polynomials of degree $n-1$ and $n-2$, respectively.

Set $\alpha_0^{(2n-1)} = -\infty$. The proof hinges on the following three claims:

- the sign of $S_{2n-2}(x)$ alternates on $\alpha_i^{(2n-1)}$, $\alpha_{i+1}^{(2n-1)}$ for $i = 0, 1, \dots, n-3$;
- the sign of $S'_{2n-1}(x)$ alternates on $\alpha_i^{(2n-1)}$, $\alpha_{i+1}^{(2n-1)}$ for $i = 1, \dots, n-2$;
- $\operatorname{sign}(S_{2n-2}(\alpha_1^{(2n-1)})) = \operatorname{sign}(S'_{2n-1}(\alpha_1^{(2n-1)}))$.

The first claim follows from the hypotheses.

The second claim follows from the fact that $S'_{2n-1}(x)$ is a polynomial of degree $n-2$ and it has exactly one root in every interval $\left(\alpha_i^{(2n-1)}, \alpha_{i+1}^{(2n-1)}\right)$ for $i = 1, 2, \dots, n-2$, as it must have a root between consecutive roots of $S_{2n-1}(x)$.

The third claim follows from the facts that

$$\operatorname{sign}\big(S_{2n-2}\big(\alpha_1^{(2n-1)}\big)\big) = -\operatorname{sign}\big(S_{2n-2}(-\infty)\big) = -(-1)^{n-1},$$

as $S_{2n-2}(x)$ has a single root, $\beta_1^{(2n-2)}$, which is less than $\alpha_1^{(2n-1)}$; and

$$\operatorname{sign}\big(S'_{2n-1}\big(\alpha_1^{(2n-1)}\big)\big) = \operatorname{sign}\big(S'_{2n-1}(-\infty)\big) = (-1)^{n-2},$$

as $S'_{2n-1}(x)$ has no root less than $\alpha_1^{(2n-1)}$.

From the three claims and (23) it follows that the sign of $S_{2n}(x)/x$, and hence of $S_{2n}(x)$, alternates on $\alpha_i^{(2n-1)}, \alpha_{i+1}^{(2n-1)}$ for $i = 1, \dots, n-3$; and this fact provides the required root $\beta_{i+1}^{(2n)}$ between these numbers, $i = 1, \dots, n-3$.

From the proof of the third claim and (23) it follows that $\operatorname{sign}\big(S_{2n}(\alpha_1^{2n-1})/\alpha_1^{2n-1}\big) = (-1)^n$. If we show that $S_{2n}(x)/x$ has a different sign at $-\infty$, then we provide the required root $\beta_1^{(2n)} < \alpha_1^{(2n-1)}$ for $S_{2n}(x)/x$, and hence for $S_{2n}(x)$. Indeed, the degree of $S_{2n-2}(x)$ is greater than the degree of $S'_{2n-1}(x)$, and therefore the sign of $S_{2n}(x)/x$ at $-\infty$ is the sign of $S_{2n-2}(x)$ at $-\infty$, namely $(-1)^{n-1}$.

As $S_{2n-2}(\alpha_{n-1}^{(2n-1)}) = S_{2n-2}(0) = 0$, the second and the third claim, and (23) imply that $S_{2n}(x)/x$ alternates on $\alpha_{n-2}^{(2n-1)}, \alpha_{n-1}^{(2n-1)}$, providing the required root $\beta_{n-1}^{(2n)}$ between these numbers, also for $S_{2n}(x)$. Finally, the last root to find is $\beta_n^{(2n)} = 0$.

Next we prove **(ii)**. In our setting the identity (18) specifies to

$$\frac{S_{2n+1}(x)}{x} = 2nS_{2n-1}(x) + S'_{2n}(x), \tag{24}$$

where the RHS is the sum of two polynomials of degree $n-1$. The proof hinges on the following three claims:

- the sign of $S_{2n-1}(x)$ alternates on $\beta_i^{(2n)}, \beta_{i+1}^{(2n)}$ for $i = 1, \dots, n-2$;
- the sign of $S'_{2n}(x)$ alternates on $\beta_i^{(2n)}, \beta_{i+1}^{(2n)}$ for $i = 1, \dots, n-1$;
- $\operatorname{sign}(S_{2n-1}(\beta_1^{(2n)})) = \operatorname{sign}(S'_{2n}(\beta_1^{(2n)}))$.

The first claim follows from the hypotheses.

The second claim follows from the fact that $S'_{2n}(x)$ is a polynomial of degree $n-1$ and it has exactly one root in every interval with the endpoints $\beta_i^{(2n)}, \beta_{i+1}^{(2n)}$ for $i = 1, 2, \dots, n-1$, as it must have a root between consecutive roots of $S_{2n}(x)$.

The third claim follows from the facts that

$$\operatorname{sign}\bigl(S_{2n-1}\bigl(\beta_1^{(2n)}\bigr)\bigr) = \operatorname{sign}\bigl(S_{2n-1}(-\infty)\bigr) = (-1)^{n-1},$$

and

$$\operatorname{sign}\bigl(S'_{2n}\bigl(\beta_1^{(2n)}\bigr)\bigr) = \operatorname{sign}\bigl(S'_{(2n)}(-\infty)\bigr) = (-1)^{n-1},$$

as neither $S'_{2n}(x)$ nor $S_{2n-1}(x)$ has a root less than $\beta_1^{(2n)}$.

From the three claims and (24) it follows that the sign of $S_{2n+1}(x)/x$, and hence of $S_{2n+1}(x)$, alternates on $\beta_i^{(2n)}, \beta_{i+1}^{(2n)}$ for $i = 1, \dots, n-2$; and this fact provides the required root $\alpha_i^{(2n+1)}$ between these numbers, $i = 1, \dots, n-2$.

As $S_{2n-1}(\beta_n^{(2n)}) = S_{2n-1}(0) = 0$, the second and the third claim, and (24) imply that $S_{2n+1}(x)/x$ alternates on $\beta_{n-1}^{(2n)}, \beta_n^{(2n)}$, providing the required root $\alpha_{n-1}^{(2n+1)}$ between these numbers, also for $S_{2n}(x)$. Finally, the last root to find is $\alpha_n^{(2n+1)} = 0$.

3. Semilabeled trees and set partitions

Let $F(n,k)$ denote the number of *rooted semilabeled trees* with k uniquely labeled leaves and n non-root vertices. Such trees have a root, which may or may not have degree one, and is not being counted as vertex or leaf; and have k leaves. Two such trees are identical, if there is a graph isomorphism between them that maps root to root and every leaf label to the same leaf label.

Erdős and Székely in [7] established a bijection between the trees counted by $F(n,k)$ and partitions of an n-element set into $n-k+1$ classes, under which out-degrees of non-root vertices and the root correspond to class sizes in the partition. (Their result immediately implies that $F(n,k) = S(n, n-k+1)$.) We use the term *phylogenetic tree* for semilabeled trees, in which the root degree is $\geqslant 2$ and every internal vertex has degree $\geqslant 3$. Let $F^\star(n,k)$ denote the number of phylogenetic trees with k leaves and n non-root vertices. The bijection provides $F^\star(n,k) = S^\star(n, n-k+1)$ and $S^\star(n,i) = F^\star(n, n-i+1)$.

Any information available on the S $(S^\star)$ numbers kind can translate to information on the F $(F^\star)$ numbers. The central and local limit theorems for $S(n,k)$ [12] translate into such for $F(n,k)$ (with $\mathcal{E}(F(n,.)) = n+1-\mathcal{E}(S(n,.))$ and $\mathcal{D}(F(n,.)) = \mathcal{D}(S(n,.))$); the central and local limit theorems for $S^\star(n,k)$ (Theorem 1) translate into such for $F^\star(n,k)$ (with $\mathcal{E}(F^\star(n,.)) = n+1-\mathcal{E}(S^\star(n,.))$ and $\mathcal{D}(F^\star(n,.)) = \mathcal{D}(S^\star(n,.))$).

Felsenstein [9, 10], and also Foulds and Robinson [11] investigated the numbers $T_{n,m}$. $T_{n,m}$ is the number of rooted trees with n labeled leaves, m unlabeled internal vertices (the root is one of them), where the root has degree at least 2 and no other internal vertices have degree 2. Clearly,

$$T_{n,m} = F^\star(n+m-1, n) = S^\star(n+m-1, m). \tag{25}$$

If we are interested only in evaluating certain $T_{n,m}$ numbers, formula (25) would suffice. However, the $T_{n,m}$ notation suggests that the distributions of $F(n,k)$ and $F^\star(n,k)$ for a large but fixed number of vertices n and a varying number of leaves k, being mathematically interesting, is not however relevant for phylogenetics. The distribution relevant for phylogenetics corresponds to large but fixed number of leaves n, and varying number of internal vertices, with which the total number of vertices varies as well. Let $t_n = \sum_k T_{n,k}$ denote a number of all phylogenetic trees with n labeled leaves. This sequence is A000311 in [17], which is the solution to Schroeder's fourth problem [16]. Next we prove central and local limit theorems for the array $T_{n,k}$.

4. Phylogenetic trees and set partitions in another distribution

Theorem 2. *For the array $A(n,j) = T_{n+1,j}$, the central limit theorem* (2) *and the local limit theorem* (5) *hold with*

$$\mathcal{E}(T_{n+1,\cdot}) = \frac{1-\rho}{2\rho}n + \frac{\frac{3}{4}-\ln 2}{\rho} + O\left(\frac{1}{n}\right)$$

and

$$\mathcal{D}^2(T_{n+1,\cdot}) = \frac{n}{4}\left(\frac{1}{\rho^2} - \frac{2}{\rho} - 1\right) + \frac{1+4\ln 2 - 8\ln^2 2}{8\rho^2} + O\left(\frac{1}{n}\right),$$

where $\rho = -1+2\ln 2$. *Furthermore, the number* $k = J_n$ *that maximizes* $T_{n+1,k}$ *satisfies*

$$J_n = \frac{1-\rho}{2\rho}n + o(\sqrt{n}), \tag{26}$$

and

$$T_{n+1,J_n} = \frac{n!(1+o(1))}{\pi\sqrt{2}n\rho^{n+1/2}\sqrt{\left(\frac{1}{\rho^2} - \frac{2}{\rho} - 1\right)}}. \tag{27}$$

PROOF. Felsenstein [9, 10] proved the recurrence relation [2)]

$$T_{n,k} = (n+k-2)T_{n-1,k-1} + kT_{n-1,k} \tag{28}$$

for $k > 1$ with the initial condition $T_{n,1} = 1$ for $n > 1$. Consider the polynomials $P_n(x) = \sum_k T_{n+1,k}x^k$. Then $P_n(1) = t_{n+1}$ and the degree of $P_n(x)$ is n. Felsenstein's recurrence relation (28) implies the identity

$$P_n(x) = nxP_{n-1}(x) + (x+x^2)P'_{n-1}(x) \tag{29}$$

with the initial term $P_0(x) = 1$, $P_1(x) = T_{2,1}x = x$. We get for the expectation and variance, from (1), using (29) repeatedly,

$$\mathcal{E}(T_{n+1,\cdot}) = \frac{t_{n+2}}{2t_{n+1}} - \frac{n+1}{2}, \tag{30}$$

$$\mathcal{D}^2(T_{n+1,\cdot}) = \frac{t_{n+3}}{4t_{n+1}} - \frac{t_{n+2}^2}{4t_{n+1}^2} - \frac{t_{n+2}}{2t_{n+1}} - \frac{n+1}{4}. \tag{31}$$

[2)] The recurrence is based on a case analysis whether the n^{th} leaf is to be grafted into an edge or to be joined to an internal vertex of an already existing tree with $n-1$ leaves.

Consider the following bivariate generating function for $T_{n,k}$:

$$H(x,z) = \sum_{n\geqslant 1}\sum_{k} T_{n,k}x^k \frac{z^n}{n!} = \sum_{n\geqslant 1} P_{n-1}(x)\frac{z^n}{n!}, \tag{32}$$

in particular, $H(1,z) = \frac{z}{1!} + \frac{z^2}{2!} + \frac{4z^3}{3!} + \frac{26z^4}{4!} + \dots$. Flajolet [8] observed the functional equation

$$H(x,z) = z + x\big(e^{H(x,z)} - 1 - H(x,z)\big), \tag{33}$$

which immediately follows from the Exponential Formula, and obtained from this equation an expression for $H(1,z)$ in terms of the Lambert function:

$$H(1,z) = -\mathrm{LambertW}\Big(-\frac{1}{2}e^{(z-1)/2}\Big) + \frac{z-1}{2}.$$

He also observed that $H(1,z)$, the EGF of the t_n sequence, has a singularity at $\rho = -1 + 2\ln 2$, and it is the only singularity at this radius; and furthermore, for $|z| < \rho$, there is a singular expansion of $H(1,z)$ in terms of $\Delta = \sqrt{1 - z/\rho}$, of which the first few terms are

$$H(1,z) = \ln 2 - \sqrt{\rho}\Delta + \Big(\frac{1}{6} - \frac{1}{3}\ln 2\Big)\Delta^2 - \frac{\rho^{3/2}}{36}\Delta^3 + O(\Delta^4). \tag{34}$$

Flajolet [8] used (34) to obtain an asymptotic formula for t_n and noted that an asymptotic expansion can be obtained by this method. Using Maple, we went further and obtained the following asymptotic expansion:

$$t_n \sim \frac{n!}{\sqrt{\pi}\rho^{n-1/2}}\Big(\frac{1}{2n^{3/2}} + \frac{3}{16n^{5/2}} + \frac{25}{256n^{7/2}} + O\Big(\frac{1}{n^{9/2}}\Big)\Big). \tag{35}$$

Combining (30) and (31) with (35), one obtains the asymptotics for the expectation and the variance in Theorem 2. The details are on a Maple worksheet [19]. □

LEMMA 2. *For $n \geqslant 1$, the polynomial $P_n(x)$ has n distinct real roots, one of them is zero, and the other $n-1$ roots are in the open interval $(-1,0)$.*

We prove the theorem by induction on n. The small cases above are easy to verify. It is easy to see (by a different induction) that $P_1(-1) = -1$ and from (29),

$P_n(-1) = (-n)P_{n-1}(-1)$, thus

$$\operatorname{sign}(P_n(-1)) = (-1)^n. \tag{36}$$

Using the induction hypothesis, let the roots of $P_n(x)$ be $-1 < \alpha_1 < \cdots < \alpha_{n-2} < \alpha_{n-1} < \alpha_n = 0$. By Rolle's theorem, $P_n'(x)$ has a root β_i in (α_i, α_{i+1}) for $i = 1, 2, \ldots, n-1$. From (29) we observe that $\operatorname{sign}(P_{n+1}(\beta_i)) = -\operatorname{sign}(P_n(\beta_i))$. As the sign of $P_n(x)$ must alternate on the β_i, so does $P_{n+1}(x)$, and therefore $P_{n+1}(x)$ has a root in (β_i, β_{i+1}) for $i = 1, 2, \ldots, n-2$. We have to find three more roots: one is $x = 0$, and we will show that the other two are in the intervals $(-1, \beta_1)$ and $(\beta_{n-1}, 0)$, respectively.

Indeed, $\operatorname{sign}(P_n(x))$ differs in -1 and β_1, since $P_n(x)$ has a single root α_1 between. Also, $\operatorname{sign}(P_{n+1}(-1)) = - \operatorname{sign}(P_n(-1))$ by (36) and $\operatorname{sign}(P_{n+1}(\beta_1)) = -\operatorname{sign}(P_n(\beta_1))$ by our earlier observation. Hence $\operatorname{sign}(P_{n+1}(x))$ differs in -1 and β_1, and therefore, $P_{n+1}(x)$ has a root in $(-1, \beta_1)$.

Observe that (29) together with the induction hypothesis imply that for $n \geqslant 1$ the coefficient of x^n in $P_n(x)$ is positive. On one hand, for negative x sufficiently close to zero we have $\operatorname{sign}(P_{n+1}(x)) = -1$. On the other hand, $\operatorname{sign}(P_{n+1}(\beta_1)) = -\operatorname{sign}(P_{n+1}(-1)) = (-1)^n$, $\operatorname{sign}(P_{n+1}(\beta_i)) = (-1)^{n+i-1}$, and $\operatorname{sign}(P_{n+1}(\beta_{n-1})) = 1$. Therefore, $P_{n+1}(x)$ has a root in $(\beta_{n-1}, 0)$.

Acknowledgements. The first two and last two authors were supported in part by the HUBI MTKD–CT-2006–042794. The first and third authors were partially supported by a grant from the University of South Carolina Promising Investigator Research Award. Péter L. Erdős was supported in part by the Hungarian NSF contracts Nos T37846, T34702, T37758. László A. Székely was supported in part by the NSF DMS contracts No. 0701111 and 1000475, the NIH NIGMS contract 3R01GM078991–03S1, and by the Alexander von Humboldt Foundation at the Rheinische Friedrich-Wilhelms Universität, Bonn.

Bibliography

1. ***H. D. Becker,*** *Solution to problem E* 461, Amer. Math. Monthly **48** (1941), 701–702.
2. ***E. R. Canfield,*** *Central and local limit theorems for the coefficients of polynomials of binomial type*, J. Combin. Theory A **23** (1977), № 3, 275–290.
3. ***E. R. Canfield,*** *Engel's inequality for Bell numbers*, J. Comb. Theory A **72** (1995), 184–187.
4. ***E. R. Canfield,*** bellMoser.pdf, 6 pages manuscript.

5. ***E. R. Canfield, L. H. Harper,*** *A simplified guide to large antichains in the partition lattice*, Congr. Numer. **100** (1994), 81–88.
6. ***L. Clark,*** *Central and local limit theorems for excedances by conjugacy class and by derangement*, Integers **2** (2002), Paper A3, 9 pp. (electronic).
7. ***P. L. Erdős, L. A. Székely,*** *Applications of antilexicographic order I: An enumerative theory of trees*, Adv. Appl. Math. **10** (1989), 488–496.
8. ***P. Flajolet,*** *A problem in statistical classification theory*, http://algo.inria.fr/libraries/autocomb/schroeder-html/schroeder.html
9. ***J. Felsenstein,*** *The number of evolutionary trees*, Syst. Zool. **27**(1) (1978), 27–33. Corrigendum Syst. Zool. **30** (1981), 122.
10. ***J. Felsenstein,*** *Inferring Phylogenies*, Sinauer Associates, Sunderland, Massachusetts, 2004.
11. ***L. R. Foulds, R. W. Robinson,*** *Enumeration of phylogenetic trees without points of degree two*, Ars Combin. **17** (1984), A, 169–183.
12. ***L. H. Harper,*** *Stirling behaviour is asymptotically normal*, Ann. Math. Stat. **38** (1967), 410–414.
13. ***E. H. Lieb,*** *Concavity properties and a generating function for Stirling numbers*, J. Comb. Theory **5** (1968), 203–206.
14. ***L. Moser, M. Wyman,*** *An asymptotic formula for the Bell numbers*, Trans. Roy. Soc. Canada III **49** (1955), 49–53.
15. ***B. Salvy, J. Shackell,*** *Asymptotics of the Stirling numbers of the second kind*, Studies in Automatic Combinatorics II (1997). Published electronically.
16. ***E. Schroeder,*** *Vier combinatorische Probleme*, Z. f. Math. Phys. **15** (1870), 361–376.
17. ***N. J. A. Sloane,*** The On-Line Encyclopedia of Integer Sequences http://www.research.att.com/∼njas/sequences/
18. http://www.math.sc.edu/∼szekely/Aprilattemptformal.pdf
19. http://www.math.sc.edu/∼szekely/copykiserletformal.pdf

Éva Czabarka,
Virginia Johnson,
László A. Székely

University of South Carolina,
SC 29208 Columbia, USA
czabarka@math.sc.edu;
johnsonv@math.sc.edu;
szekely@math.sc.edu

Péter L. Erdős

Alfréd Rényi Institute,
13–15 Reáltanoda u.,
1053 Budapest, Hungary
elp@renyi.hu

Anne Kupczok

IST Austria
Am Campus 1
3400 Klosterneuburg, Austria
anne.kupczok@ist.ac.at