Scan A5263 etc

Fowler & Robinson (Paper)

add to many

DETERMINING THE ASYMPTOTIC NUMBER OF PHYLOGENETIC TREES

L.R. FOULDS AND R.W. ROBINSON*

The phylogenetic (evolutionary) trees of biology are a special class of labelled trees. In graph-theoretic terms, a phylogenetic tree is a tree whose points have been labelled with disjoint subsets of the labelling set. Every point of degree less than three must have a nonempty label. Formulas are found for the exact numbers of phylogenetic trees with n labels, and numerical results are presented for selected $n \leq 40$. The asymptotic behaviour of these numbers as $n \to \infty$ is determined. Similar results are obtained for the mean and variance of the number of points in a phylogenetic tree with n labels. The effect of requiring that each nonempty label be a singleton is also studied.

1. INTRODUCTION

It has been postulated that existing biological species have been linked in the past by common ancestors. A diagram showing these links is called a phylogeny or phylogenetic tree.

Mathematically we define a *phylogenetic tree* to be a tree in the ordinary graph theoretic sense together with a map from some set $\{1,2,\ldots,n\}$ of *labels* to the point set of the tree such that every point of degree less than three is in the image. It should be noted that some points may have several labels while others may have none. The number n of labels is termed the *magnitude* of the phylogenetic tree. The *order* of a tree is the number of points in it. These trees, together with their significance in molecular evolution, are discussed in Robinson and Foulds [8]. Graph theoretic terminology not defined in this paper can be found in the book by Harary [4].

A number of methods for constructing phylogenies have been proposed. Trees produced by various authors for similar sets of species are presented in Foulds *et al.* [2]. If phylogenies for a given set of n species are to be generated, it is of interest to know how many possible trees exist. These numbers are calculated exactly for given n by means of recurrences derived in the next section of this paper. The mean and variance of the number of points in the trees with magnitude n are also found. The asymptotic behaviour of the exact numbers and the statistics are determined in Section 3 using the methods of Harary *et al.* [3]. A similar analysis is then carried out in Section 4 with the restriction that all label sets must be singletons. Tables of numerical results for selected values of n up to 40 are presented in Section 5,

---

* The second author is grateful for the support of the Australian Research Grants Committee for the project "Numerical Implementation of Unlabeled Graph Counting Algorithms", under which research and computing for this paper were performed.

including a comparison of exact values with the asymptotic estimates obtained earlier.

2. COUNTING PHYLOGENETIC TREES

Let T_n be the number of different phylogenetic trees of magnitude n. Our object in this section is to derive recurrence relations by which T_n can be calculated for successive values of n. The exponential generating function defined by

$$T(x) = \sum_{n=1}^{\infty} T_n x^n/n!$$

will be a useful tool in our analysis. In order to establish the mean and variance of the numbers of points in phylogenies of given magnitude we shall also find recurrences involving the number $T_{n,p}$ of different phylogenies with magnitude n and order p. The corresponding generating function is given by

$$T(x,y) = \sum_{n=1}^{\infty} \sum_{p=1}^{2n-2} T_{n,p} x^n y^p/n! \ .$$

Thus T(x) is obtained from T(x,y) by setting y to 1. The fact that $1 \leq p \leq 2n-2$ is shown in [8].

As usual in tree counting, the numbers are first determined for trees which have some point distinguished as the *root*. In particular, let a *planted phylogeny* be a tree rooted at an endpoint and labelled according to the rules for an ordinary phylogeny except that the root is not to receive any label. Likewise, the root point of a planted phylogeny is not counted in its order. In the biological context, a planted phylogeny corresponds to a phylogenetic tree in which a common ancestor is designated. This is represented diagramatically by orienting all lines away from the point representing the common ancestor.

Let P_n denote the number of different planted phylogenies of magnitude n, and let $P_{n,p}$ denote the number of these of order p. The associated generating functions are

$$P(x) = \sum_{n=1}^{\infty} P_n x^n/n! \ ,$$

$$P(x,y) = \sum_{n=1}^{\infty} \sum_{p=1}^{2n-1} P_{n,p} x^n y^p/n! \ .$$

For a fixed magnitude $n \geq 1$ there is just one planted phylogeny of order 1, which is termed the *trivial* phylogeny. Thus, the exponential generating function for trivial planted phylogenies is

$$(x + \frac{x^2}{2} + \frac{x^3}{6} + \ldots)y = (e^x - 1)y \ .$$

Any non-trivial planted phylogeny can be viewed as the result of joining one or more planted phylogenies at their roots, these being identified as a single ordinary point which is then joined to a new root. In this process the original root point must receive some label if it becomes a point of degree 2, and in any case may receive new labels. Thus the generating function for planted phylogenies in which the point adjacent to the root has degree 2 is

$$(e^x - 1)yP(x,y).$$

Here, as above, $(e^x - 1)y$ enumerates the possibilities for the point adjacent to the root, while $P(x,y)$ enumerates the possibilities for completing the tree. As usual in labelled counting problems the product of the exponential generating function accounts for the number of ways in which the sets of labels from the two parts can be obtained from the label set of the union. (See for example Chapter 1 of Harary and Palmer [5].) Similarly, if the point adjacent to the root has degree $k+1 \geq 3$, the number of possibilities is enumerated by

$$e^x y \ P(x,y)^k/k! \ .$$

In this expression $e^x y$ accounts for the point adjacent to the root, since it need not be labelled. There are k planted phylogenies to be joined at this point, and we divide by k! because the sequence in which they are added is immaterial. Summing over $k \geq 2$, adding in the other two terms and taking advantage of the exponential form of the sum, we have

$$P(x,y) = ye^{x+P(x,y)} - y - yP(x,y) \ . \qquad (2.1)$$

Setting $y = 1$ gives

$$1 + 2P(x) = e^{x+P(x)} \ .$$

Then differentiating and simplifying to eliminate the exponential yields

$$P'(x) = 1 + 2P(x) + \left(P(x)^2\right)' \ . \qquad (2.2)$$

Given that $P_0 = 0$, comparing coefficients of $x^{n-1}/(n-1)!$ in this equation gives $P_1 = 1$ and

$$P_n = 2P_{n-1} + \sum_{0<i<n} P_i P_{n-i} \binom{n}{i} \qquad (2.3)$$

for $n \geq 2$.

The generating function for the unrooted phylogenetic trees $T(x,y)$ can now be determined in terms of phylogenetic trees in which a point is distinguished (*point-rooted*) and those in which a line is distinguished (*line-rooted*). A point-rooted tree is associated in a 1-1 fashion with the planted tree obtained by joining a new root point to the original. All planted trees are obtained in this way except for those in which the point adjacent to the root has degree 3 and has received no label. The latter are counted by the exponential generating function $yP(x,y)^2/2$, and so point-rooted trees are counted by

$$P(x,y) - yP(x,y)^2/2 \quad .$$

A line-rooted tree can be viewed as the result of identifying the roots of two planted trees, which are then suppressed to form the root line. These are therefore enumerated by

$$P(x,y)^2/2 \ ,$$

the factor of 2 accounting for the fact that the same line-rooted tree is obtained by interchanging the two planted trees.

The difference between our expressions for point- and line-rooted trees will be $T(x,y)$. For on assembling the terms corresponding to a particular unrooted tree of magnitude n and order p we find $px^n y^p/n! - (p-1)x^n y^p/n! = x^n y^p/n!$. This is because the tree has no automorphisms, having been labelled at all its endpoints, so that all p points give distinct point-rooted trees while all p-1 lines give distinct line-rooted trees. Thus we have

$$T(x,y) = P(x,y) - \frac{y+1}{2} P(x,y)^2 \ . \qquad (2.4)$$

Setting $y = 1$ we obtain

$$T(x) = P(x) - P(x)^2 \ . \qquad (2.5)$$

On differentiating with respect to x and comparing with (2.2) one sees that

$$T'(x) = 1 + 2P(x).$$

We can compare coefficients of $x^{n-1}/(n-1)!$ to determine that $T_1 = 1$ and

$$T_n = 2P_{n-1} \tag{2.6}$$

for $n \geq 2$. Thus (2.3) can be used to compute values of T_n. The results for selected n up to 40 are listed in Section 5.

Since a phylogenetic tree of magnitude n may have any order from 1 to 2n-2, it is of interest to determine the mean μ_n and standard deviation σ_n of the number of points among all such trees. This will be done by finding recurrence relations for the first and second moments about the origin, namely

$$T_n^{(i)} = \sum_{p=1}^{2n-2} p^i T_{n,p}, \qquad i = 1 \text{ or } 2.$$

Then as usual we have

$$\mu_n = T_n^{(1)}/T_n \ ,$$

$$\sigma_n^2 = (T_n^{(2)}/T_n) - \mu_n^2 \ .$$

The exponential generating functions for the moments are

$$T^{(i)}(x) = \sum_{n=1}^{\infty} T_n^{(i)} x^n/n!, \qquad i = 1 \text{ or } 2 \ .$$

Since

$$T^{(1)}(x) = T_y(x,1)$$

and

$$T^{(2)}(x) = T_{yy}(x,1) + T_y(x,1) \ , \tag{2.7}$$

we can evaluate the moment generating functions on the basis of previous equations. Differentiating (2.4) gives

$$T_y(x,y) = P_y(x,y) - \tfrac{1}{2}P(x,y)^2 - (y+1)P(x,y)P_y(x,y) \tag{2.8}$$

Now differentiating (2.1) we find

$$P_y(x,y) = \frac{P(x,y)}{y} + (y+1)P(x,y)P_y(x,y) \qquad (2.9)$$

which with (2.8) simplifies to

$$T_y(x,y) = \frac{P(x,y)}{y} - \tfrac{1}{2}P(x,y)^2 \quad .$$

Differentiating again and simplifying with (2.9) yields

$$T_{yy}(x,y) = P(x,y)^2/y^2\left[1-(1+y)P(x,y)\right] \quad .$$

Setting $y = 1$ in both equations and using (2.7), we have

$$T^{(1)}(x) = P(x) - \tfrac{1}{2}P(x)^2 \quad , \qquad (2.10)$$

$$T^{(2)}(x) = T^{(1)}(x) + P(x)^2/\left[1-2P(x)\right] \quad .$$

Recurrence relations are obtained by comparing coefficients of $x^n/n!$. The first of these can be simplified using (2.3), giving $T^{(1)} = 1$ and

$$T_n^{(1)} = \tfrac{1}{2}P_n + P_{n-1} \qquad (2.11)$$

for $n \geq 2$. In the second equation we write

$$T_n^{(2)} = T_n^{(1)} + S_n \quad , \qquad (2.12)$$

where S_n is the contribution of the right hand term. Thus $S_0 = 0$, $S_1 = 0$, and for $n \geq 2$, S_n is determined from P_i $(1 \leq i \leq n)$ by the recurrence

$$S_n = P_n - 2P_{n-1} + 2 \sum_{0<i<n} P_{n-i} S_i \binom{n}{i} \quad . \qquad (2.13)$$

Now from the values of P_n and T_n mentioned earlier and the last three relations one can compute values of $T_n^{(1)}$ and $T_n^{(2)}$, and hence of μ_n and σ_n. The results for $1 \leq n \leq 10$ and n up to 40 by 5's are listed in Section 5.

3. ASYMPTOTIC BEHAVIOUR

In this section the asymptotic behaviour of T_n, μ_n and σ_n as $n \to \infty$ is determined. Since P_n is the basis for the equations defining these quantities, we start with a study of the exponential generating function $P(x)$. Setting y to 1 in (2.1) we had

$$1 + 2P(x) = e^{x+P(x)}, \tag{3.1}$$

from which we can solve for x to obtain

$$x = \log\big(1+2P(x)\big) - P(x) \ . \tag{3.2}$$

Regarding P as a complex variable, x is clearly analytic for $P \neq -\frac{1}{2}$, and the derivative $(1-2P)/(1+2P)$ is nonzero when $P \neq \frac{1}{2}$. Also $x = 0$ when $P = 0$, so the inverse function is analytic in some neighbourhood of $x = 0$. The power series expansion of this function about $x = 0$ is our generating function $P(x)$, which is now seen to have a positive radius of convergence, say ρ. As $P_n \geq 1$ for $n \geq 1$ we see that $\rho < \infty$, since $P(x)$ cannot attain the value $\frac{1}{2}$ within the circle of convergence. Then $x = \rho$ is a singularity of $P(x)$ by Pringsheim's Theorem (see Hille [6, p.133]), and since $P(\rho) > 0$ we must have $P(\rho) = \frac{1}{2}$. Also, for $|x| = \rho$ and $x \neq \rho$ we have $|P(x)| < \frac{1}{2}$ so that $x = \rho$ is the sole singularity of $P(x)$ on its circle of convergence. Finally, setting $x = \rho$ in (3.2) gives

$$\rho = \ln 2 - \tfrac{1}{2} \ . \tag{3.3}$$

So far we have established the results required of steps 1-11 in the 20 step algorithm of [3], by methods which are more direct than usual for tree counting problems. However the remainder of the development is standard, and so we refer to this paper for the explanations of the remaining steps and confine ourselves to performing the necessary calculations.

At $P = \frac{1}{2}$ the second derivative of x as a function of P is -1, and so from steps 12 and 13 we have that $x = \rho$ is a branch point of order 2 for $P(x)$. Thus as in step 14 one has an expansion of the form

$$P(x) = \tfrac{1}{2} - b_1(\rho-x)^{1/2} + b_2(\rho-x) + b_3(\rho-x)^{3/2} + \ldots$$

valid in some neighbourhood of $x = \rho$. Substituting into equation (2.2) gives a relation which must be satisfied by this expression. One can then compare coefficients of $(\rho-x)^{-\frac{1}{2}}, (\rho-x)^{0}, (\rho-x)^{\frac{1}{2}}, \ldots$ to determine as many of $b_1, b_2, \ldots$ as required. In

particular we find $b_1^2 = 2$, so that $b_1 = \sqrt{2}$ in order for the expansion around 0 to agree with the expansion around ρ where their circles of convergence overlap. Two more comparisons then establish that $b_2 = 2/3$ and $b_3 = -1/9\sqrt{2}$, so that

$$P(x) = \tfrac{1}{2} - 2^{1/2}(\rho-x)^{1/2} + \tfrac{2}{3}(\rho-x) - \tfrac{1}{9}2^{-1/2}(\rho-x)^{3/2} \pm \dots \qquad (3.4)$$

To evaluate the contribution of a term $(\rho-x)^{k/2}$ in this expansion, note that the coefficient of x^n in $(1-x)^{-s}$ is just $\Gamma(s+n)/\Gamma(s)\Gamma(n+1)$ provided s is not a non-negative integer. From Stirling's formula the latter is

$$\frac{n^{s-1}}{\Gamma(s)}\left(1 + \frac{s(s-1)}{2n} + O(\tfrac{1}{n^2})\right) \qquad (3.5)$$

as $n \to \infty$. Thus the term $-b_1(\rho-x)^{\frac{1}{2}}$ contributes

$$(2\pi)^{-1/2}\rho^{1/2}n^{-3/2}\rho^{-n}\left(1 + \frac{3}{8n} + O(\tfrac{1}{n^2})\right) .$$

The next term, $b_3(\rho-x)^{3/2}$, contributes

$$-\frac{1}{12}(2\pi)^{-1/2}\rho^{3/2}n^{-5/2}\rho^{-n}\left(1 + O(\tfrac{1}{n})\right)$$

when taken to the same order. The remaining terms collectively contribute $O(n^{-7/2}\rho^{-n})$, as can be seen from Darboux's Theorem (as in Theorem 4 of Bender [1]) or from Pólya's Lemma (as in [3]). In sum we have

$$\frac{P_n}{n!} = \left(\frac{\rho}{2\pi}\right)^{1/2} n^{-3/2}\rho^{-n}\left(1 + \frac{5-\ln 2}{12n} + O(\tfrac{1}{n^2})\right) , \qquad (3.6)$$

in view of $\rho = \ln 2 - 1/2$.

Since $T_n = 2P_{n-1}$ and $T_n^{(1)} = \frac{1}{2}P_n + P_{n-1}$ for $n > 1$ we also have

$$\frac{T_n}{n!} = 2\rho\left(\frac{\rho}{2\pi}\right)^{1/2} n^{-5/2}\rho^{-n}\left(1 + \frac{23 - \ln 2}{12n} + O(\tfrac{1}{n^2})\right) , \qquad (3.7)$$

and

$$\frac{T_n^{(1)}}{n!} = \tfrac{1}{2}\left(\frac{\rho}{2\pi}\right)^{1/2} n^{-3/2}\rho^{-n}\left(1 + \frac{23\ln 2 - 7}{12n} + O(\tfrac{1}{n^2})\right) . \qquad (3.8)$$

Taking the ratio of (3.8) to (3.7) gives the average number of points in a tree of magnitude n as

$$\mu_n = \frac{n}{4\rho}\left(1 + \frac{2\ln 2 - 5/2}{n} + O(\frac{1}{n^2})\right) \quad . \tag{3.9}$$

To analyse the variance similarly, start with the exponential generating function $S(x) = P(x)^2/1-2P(x)$. Substituting the expansion (3.4) for $P(x)$ yields

$$S(x) = 2^{-7/2}(\rho-x)^{-1/2} - \frac{11}{24} + \frac{49}{3}\,2^{-9/2}(\rho-x)^{1/2} + \ldots \quad . \tag{3.10}$$

As before, (3.5) can be applied in conjunction with Darboux's Theorem or Pólya's Lemma to evaluate the coefficients of $S(x)$ asymptotically. The result is

$$\frac{S_n}{n!} = 2^{-7/2}(\pi\rho)^{-1/2}n^{-1/2}\rho^{-n}\left(1 + \frac{138 - 294\ln 2}{72n} + O(\frac{1}{n^2})\right) \quad .$$

Finally $\sigma_n^2 = (S_n/T_n) + \mu_n - \mu_n^2$, so this can be combined with (3.7) and (3.9), giving

$$\sigma_n^2 = \frac{1-4\rho}{16\rho^2}\cdot n + O(1) \quad . \tag{3.11}$$

In computing the variance the leading terms added out, with the effect that while the average is $O(n)$, the standard deviation is $O(n^{1/2})$. Thus the distribution of the number of points in a tree of weight n becomes relatively more sharply peaked with increasing n.

4. SINGLETON LABELS

Let $\overline{T}_n$ denote the number of phylogenies of magnitude n in which every non-empty set of labels contains just one member. Our object is to evaluate $\overline{T}_n$ exactly and asymptotically, and to do the same for the average and variance of the point distributions for these trees. The obvious approach would be to parallel almost exactly the development outlined in the previous sections, with appropriate changes in certain details. However there is a very simple relation between T_n and $\overline{T}_n$ which will enable the same results to be obtained much more quickly.

Once again the exponential generating function is a natural tool. We let

$$\overline{T}(x) = \sum_{n=1}^{\infty} \overline{T}_n x^n/n! \quad .$$

Every tree counted by $T(x)$ can be obtained from a tree counted by $\overline{T}(x)$ by enlarging the set of labels. Each singleton label set can be replaced by a set with $1,2,3,\ldots$ labels. The exponential generating function for these possibilities is $e^x-1 = x + x^2/2! + x^3/3! + \ldots$, since among themselves the k labels of a particular set have just one ordering. Phylogenies with exactly n non-empty sets of labels are counted by $\overline{T}_n(e^x-1)^n/n!$, since multiplication of exponential generating functions accounts for the number of ways that the various label sets can be interleaved. Then summing over n gives

$$T(x) = \overline{T}(e^x-1) \ ,$$

which can be inverted to the form

$$\overline{T}(x) = T\big(\ln(1+x)\big) \ . \tag{4.1}$$

By the same reasoning, the replacement of x by $\ln(1+x)$ serves to transform $P(x), T^{(1)}(x)$, and $T^{(2)}(x)$ to exponential generating functions $\overline{P}(x)$, $\overline{T}^{(1)}(x)$, and $\overline{T}^{(2)}(x)$ for the singleton label set cases of planted phylogenies and the first two moments of the point distribution of free phylogenies. The transformation can be couched in terms of the Stirling numbers $s(n,k)$ of the first kind, in view of the well-known fact that

$$\sum_{n=k}^{\infty} \frac{s(n,k)x^n}{n!} = \frac{\ln(1+x)^k}{k!} \ .$$

Thus (4.1) is converted to

$$\overline{T}_n = \sum_{k=1}^{n} T_k s(n,k) \ . \tag{4.2}$$

The same applies to $\overline{T}_n^{(1)}$ and $\overline{T}_n^{(2)}$, and so the exact numbers in the singleton label set case are readily calculated from the numbers in the unrestricted case.

In order to analyse these numbers asymptotically, the radius of convergence of $\overline{P}(x)$ and its behaviour on the circle of convergence must be established. Replacing x by $\ln(1+x)$ in (3.1) one finds

$$1 + x = \big(1+2\overline{P}(x)\big)e^{-\overline{P}(x)} \ . \tag{4.3}$$

Thus x is analytic as a function of $\overline{P}$, with $x = 0$ when $\overline{P} = 0$. Differentiating, we have

$$\frac{dx}{d\overline{P}} = (1-2\overline{P})e^{-\overline{P}} ,$$

which is nonzero just when $\overline{P} \neq 1/2$. Also $\overline{P}_n \geq 1$ for all $n \geq 1$, so it follows as in the previous section that $\overline{P}(x)$ is analytic in a neighbourhood of $x = 0$ with radius of convergence σ, that $0 < \sigma < \infty$, that $x = \sigma$ is the sole singularity of $\overline{P}(x)$ on its circle of convergence, and that $\overline{P}(\sigma) = 1/2$. From the latter result and (4.3) we find that

$$\sigma = 2e^{-1/2} - 1 \quad . \tag{4.4}$$

Equations giving $\overline{T}(x)$, $\overline{T}^{(1)}(x)$ and $\overline{T}^{(2)}(x)$ in terms of $\overline{P}(x)$ are obtained from (2.5) and (2.10) when x is replaced by $\ln(1+x)$. In this way it can be seen that each of these generating functions has $x = \sigma$ as the only possible singularity in the circle $|x| \leq \sigma$. Near the singularity we replace $\rho-x$ with $\rho-\ln(1+x)$. Now $1+\sigma = e^{\rho}$, so

$$\rho - \ln(1+x) = -\ln(1 - \frac{\sigma-x}{1+\sigma}) \quad .$$

Expanding on the right in powers of $\sigma-x$ and taking the square root, we have

$$\big(\rho - \ln(1+x)\big)^{1/2} = \big(\frac{\sigma-x}{1+\sigma}\big)^{1/2} + \frac{1}{4}\big(\frac{\sigma-x}{1+\sigma}\big)^{3/2} + \ldots \quad . \tag{4.5}$$

Together with the expansion (3.4) for $P(x)$ this gives

$$\overline{P}(x) = \tfrac{1}{2} - 2^{1/2}\big(\frac{\sigma-x}{1+\sigma}\big)^{1/2} - 2^{-3/2}\big(\frac{\sigma-x}{1+\sigma}\big)^{3/2} \pm \ldots \tag{4.6}$$

for x in some neighbourhood of σ.

To find a similar expansion for $\overline{T}(x)$ efficiently requires an analogue of (2.6). Differentiating (4.1) gives

$$\overline{T}'(x) = T'\big(\ln(1+x)\big)/(1+x).$$

Replacing x by $\ln(1+x)$ in the equation preceding (2.6), we find

$$T'\big(\ln(1+x)\big) = 1 + 2\overline{P}(x) \quad .$$

Together they produce

$$\overline{T}'(x) = \big(1 + 2\overline{P}(x)\big)/(1+x) \quad , \tag{4.7}$$

which is the desired equation. Now, writing $1 + x = (1+\sigma)\big(1 - (\sigma-x)/(1+\sigma)\big)$ and using (4.6) we have

$$\overline{T}'(x) = e^{1/2} - e^{3/4}(\sigma-x)^{1/2} + \tfrac{5}{6}e(\sigma-x) - \tfrac{47}{72}e^{5/4}(\sigma-x)^{3/2} \pm \ldots \tag{4.8}$$

for x near σ.

From (2.10) we have

$$\begin{aligned} \overline{T}^{(1)}(x) &= \overline{P}(x) - \tfrac{1}{2}\overline{P}(x)^2 \quad , \\ \overline{S}(x) \quad &= \overline{P}(x)^2/\big(1 - 2\overline{P}(x)\big) \quad , \end{aligned} \tag{4.9}$$

where $\overline{S}(x) = \overline{T}^{(2)}(x) - \overline{T}^{(1)}(x)$.

With (4.6) this implies

$$\begin{aligned} \overline{T}^{(1)}(x) &= \tfrac{3}{8} - \tfrac{1}{2}e^{1/4}(\sigma-x)^{1/2} - \tfrac{1}{3}e^{1/2}(\sigma-x) + \tfrac{37}{144}e^{3/4}(\sigma-x)^{3/2} \pm \ldots \quad , \\ \overline{S}(x) \quad &= \tfrac{1}{8}e^{-1/4}(\sigma-x)^{-1/2} - \tfrac{11}{24} + \tfrac{95}{192}e^{1/4}(\sigma-x)^{1/2} \pm \ldots \quad . \end{aligned} \tag{4.10}$$

Now, as in the previous section relation (3.5) can be applied to these expansions in $(\sigma-x)^{1/2}$. This gives

$$\begin{aligned} \frac{\overline{P}_n}{n!} &= \tfrac{1}{2}e^{1/4}\pi^{-1/2}\sigma^{1/2}n^{-3/2}\sigma^{-n}\Big(1 + \frac{11e^{1/2}-4}{48n} + O(\tfrac{1}{n^2})\Big) \quad , \\ \frac{\overline{T}_n}{n!} &= \tfrac{1}{2}e^{3/4}\pi^{-1/2}\sigma^{3/2}n^{-5/2}\sigma^{-n}\Big(1 + \frac{47e^{1/2}-4}{48n} + O(\tfrac{1}{n^2})\Big) \quad , \\ \overline{\mu}_n &= \frac{n}{2\sigma}\, e^{-1/2}\Big(1 + \frac{8-7e^{1/2}}{4n} + O(\tfrac{1}{n^2})\Big) \quad , \end{aligned} \tag{4.11}$$

and

$$\overline{\sigma}_n^2 = \frac{n}{8e\sigma^2}\,(5e^{1/2} - 8)\big(1 + O(\tfrac{1}{n})\big) \quad . \tag{4.12}$$

As before $\overline{\mu}_n$ denotes the average number of points in a phylogeny of magnitude n having singleton label sets, and $\overline{\sigma}_n$ is the standard deviation of this distribution. Again

5172
5640
5264
5263

$\overline{\mu}_n = O(n)$ and $\overline{\sigma}_n = O(n^{1/2})$, so the distribution of the number of points becomes narrower as n increases.

5. NUMERICAL RESULTS

The values of P_n, T_n, $\overline{P}_n$ and $\overline{T}_n$ for $1 \leq n \leq 10$ and n = 15,20,25,30,35 and 40 are presented in Table 1. The full range of values for $1 \leq n \leq 40$ is available from the second author. Computation of P_n and T_n were based on (2.3) and (2.6). Then $\overline{T}_n$ and $\overline{P}_n$ were obtained by applying (4.2) and its analogue for the planted trees. In Table 2 the corresponding values of μ_n, σ_n^2, $\overline{\mu}_n$ and $\overline{\sigma}_n^2$ are given. This required $T_n^{(1)}$ and S_n, which were determined from (2.11) and (2.13), and $\overline{T}_n^{(1)}$ and $\overline{S}_n$, which were then determined using Stirling numbers as in (4.2). The computations were programmed on a PDP 11/45 by A. Nymeyer while employed under an A.R.G.C. grant.

n	P_n T_n $\overline{P}_n$ $\overline{T}_n$
1	1 1 1 1
2	4 2 3 1
3	32 8 22 4
4	416 64 262 32
5	7552 832 4336 396
6	1 76128 15104 91984 6692
7	50 18624 3 52256 23 81408 1 43816

Table 1. Exact Numbers of Phylogenies

n	P_n T_n $\overline{P}_n$ $\overline{T}_n$
8	1689 68192 100 37248 728 00928 37 56104
9	65632 82944 3379 36384 25666 06784 1155 53024
10	28 89091 31776 1 31265 65888 10 25152 01984 40932 36352
15	2 08198 89496 01901 60896 5956 91776 27763 67104 45349 58105 28699 24352 1155 63646 85240 67328
20	93018 62240 00428 24880 71556 95616 1942 20449 17978 87651 06001 67424 12420 03507 66670 45650 11644 04736 232 32788 43785 27949 56315 70944
25	1 57284 43230 10411 26702 99480 89652 15960 43264 2585 37191 04491 56575 28193 77679 68528 13824 12866 79332 08929 57972 68733 16671 02077 41952 190 10306 66913 48057 68985 15093 48802 23232
30	7 59321 59246 03104 19251 95545 94129 56086 14450 57748 66432 10291 77634 70290 83101 81616 82246 65625 56728 22435 67616 38048 13710 05887 64417 78763 90069 07733 87235 68876 25728 464 50699 83318 12423 68341 05995 95743 73022 72743 17824
35	87 17281 46232 86686 06823 23779 47767 41329 89579 47614 90019 03637 79072 1 00519 01170 77685 23635 06696 09822 93259 04307 88912 77330 29272 28928 2 67516 97679 88151 87936 51523 29706 91141 56231 24276 92066 76490 81344 2782 57165 32537 80110 98592 94572 67467 11087 87167 24515 55351 59296
40	2093 14161 45446 15020 97212 45396 21891 13027 41323 87988 37211 16645 09080 98849 66912 21 00155 29520 59047 86233 43248 35161 15962 58872 34302 21145 78442 80356 15416 64768 39 33638 53022 00922 81963 78448 50186 24000 33946 96336 89340 63648 03382 88097 36192 35640 23605 11776 19860 19442 08827 46565 50471 41246 74673 65971 60286 21893 14048

Table 1 (concluded)

n	μ_n	σ_n^2	$\bar{\mu}_n$	$\bar{\sigma}_n^2$
1	1·00000	0·00000	1·00000	0·00000
2	1·50000	0·25000	2·00000	0·00000
3	2·50000	0·75000	3·25000	0·18750
4	3·75000	1·18750	4·59375	0·42871
5	5·03846	1·57544	5·97475	0·68118
6	6·33051	1·95856	7·37268	0·93493
7	7·62355	2·34084	8·77936	1·18787
8	8·91706	2·72267	10·19102	1·43975
9	10·21083	3·10424	11·60575	1·69071
10	11·50475	3·48565	13·02251	1·94093
15	17·97539	5·39172	20·12104	3·18541
20	24·44666	7·29715	27·22954	4·42462
25	30·91814	9·20237	34·34162	5·66168
30	37·38972	11·10750	41·45540	6·89765
35	43·86136	13·01257	48·57010	8·13301
40	50·33302	14·91762	55·68536	9·36799

Table 2. Mean and Variance of Number of Points in Phylogenies

For each series of exact values in Tables 1 and 2 we have asymptotic estimates provided by the first two terms on the right side of (3.6), (3.7), (3.9), and (4.11), and by the first term of (3.11) and (4.12). We denote the estimate by a $\hat{}$, so that $\hat{\sigma}_n^2 = \left((1-4\rho)/16\rho^2\right)n$ and so on. In Table 3 the exact values are compared with the asymptotic estimates at n = 10,20,30 and 40. This is sufficient to indicate the relative magnitude of the errors in the approximations, and to reveal that these errors already show the expected dependence on n. For instance $(P_n-\hat{P}_n)/P_n$ gives the proportion by which the estimate is less than the actual value of P_n. This reaches almost 10^{-4} at n = 40. The relative error is $O(n^2)$, so multiplying by n^2 gives a

quantity which is O(1). Indeed it seems to be increasing very slowly, as seen in the second line of the table. Similar observations can be made for T_n, $\overline{P}_n$ and $\overline{T}_n$. The averages and variances are presented in terms of absolute error, since their magnitudes are O(n) with constants reasonably close to unity.

n	10	20	30	40
$(P_n-\hat{P}_n)/P_n$	$1{\cdot}6424\times10^{-3}$	$4{\cdot}1226\times10^{-4}$	$1{\cdot}8339\times10^{-4}$	$1{\cdot}0320\times10^{-4}$
$n^2(P_n-\hat{P}_n)/P_n$	0·16424	0·16490	0·16505	0·16512
$(T_n-\hat{T}_n)/T_n$	$2{\cdot}827\times10^{-2}$	$7{\cdot}208\times10^{-3}$	$3{\cdot}224\times10^{-3}$	$1{\cdot}819\times10^{-3}$
$n^2(T_n-\hat{T}_n)/T_n$	2·827	2·883	2·902	2·911
$(\overline{P}_n-\hat{\overline{P}}_n)/\overline{P}_n$	$5{\cdot}64\times10^{-4}$	$1{\cdot}52\times10^{-4}$	$6{\cdot}92\times10^{-4}$	$3{\cdot}94\times10^{-5}$
$n^2(\overline{P}_n-\hat{\overline{P}}_n)/\overline{P}_n$	$5{\cdot}64\times10^{-2}$	$6{\cdot}09\times10^{-2}$	$6{\cdot}23\times10^{-2}$	$6{\cdot}30\times10^{-2}$
$(\overline{T}_n-\hat{\overline{T}}_n)/\overline{T}_n$	$1{\cdot}510\times10^{-2}$	$3{\cdot}990\times10^{-3}$	$1{\cdot}803\times10^{-3}$	$1{\cdot}023\times10^{-3}$
$n^2(\overline{T}_n-\hat{\overline{T}}_n)/\overline{T}_n$	1·510	1·596	1·623	1·636
$\mu_n-\hat{\mu}_n$	$2{\cdot}78\times10^{-3}$	$1{\cdot}19\times10^{-3}$	$7{\cdot}58\times10^{-4}$	$5{\cdot}56\times10^{-4}$
$n(\mu_n-\hat{\mu}_n)$	$2{\cdot}78\times10^{-2}$	$2{\cdot}38\times10^{-2}$	$2{\cdot}27\times10^{-2}$	$2{\cdot}22\times10^{-2}$
$\hat{\sigma}_n^2-\sigma_n^2$	0·3243	0·3227	0·3222	0·3220
$\overline{\mu}_n-\hat{\overline{\mu}}_n$	$4{\cdot}89\times10^{-2}$	$2{\cdot}22\times10^{-2}$	$1{\cdot}43\times10^{-2}$	$1{\cdot}06\times10^{-2}$
$n(\overline{\mu}_n-\hat{\overline{\mu}}_n)$	0·489	0·443	0·430	0·423
$\hat{\overline{\sigma}}_n^{\,2}-\overline{\sigma}_n^{\,2}$	0·527	0·511	0·505	0·503

Table 3. Accuracy of Asymptotic Estimates

6. RELATED RESULTS

The methods of the present paper have been applied to other classes of trees which are relevant to the formation of phylogenetic diagrams in biology. These classes are determined by applying certain combinations of the following conditions: no points of degree 2 are allowed; every point has degree 1 or 3; only endpoints are labelled; each label set is a singleton. It is planned to present the results elsewhere.

The phenomenon of relatively small variance has been observed in the statistics of some other variables in labelled trees. For example when $k > 0$ is fixed the number of points of degree k in a random labelled tree has average and variance both $O(n)$; see Moon [7, p.73]. The same is also true for the number of paths of length k in a random labelled tree (see [7, p.78]).

REFERENCES

[1] E.A. Bender, Asymptotic methods in enumeration, *SIAM Rev.* 16 (1974), 485-515.

[2] L.R. Foulds, M.D. Hendy and David Penny, A graph-theoretic approach to the construction of minimal phylogenetic trees, *J. Mol. Evol.* 13 (1979), 127-150.

[3] F. Harary, R.W. Robinson and A.J. Schwenk, Twenty step algorithm for determining the asymptotic number of trees of various species, *J. Austral. Math. Soc.* Ser. A 20 (1975), 483-503.

[4] F. Harary, *Graph Theory* (Addison-Wesley, Reading, Mass., 1969).

[5] F. Harary and E.M. Palmer, *Graphical Enumeration* (Academic Press, New York, 1973).

[6] E. Hille, *Analytic Function Theory* (Vol. 1, Ginn, Boston, 1959).

[7] J.W. Moon, *Counting Labelled Trees* (Can. Math. Congress, Montreal, 1970).

[8] D.F. Robinson and L.R. Foulds, Comparison of Labelled Trees, (to appear).

Department of Mathematics and Statistics
Massey University
Palmerston North
New Zealand

Department of Mathematics
University of Newcastle
New South Wales