A LIMITING THEOREM ON 2-COLOURED TRIVALENT TREES

A. Meir
York University, N. York, Ontario, Canada M3J 1P3

J.W. Moon
University of Alberta, Edmonton, Alberta, Canada T6G 2G1

M.S. Steel
University of Canterbury, Christchurch, New Zealand

ABSTRACT. Leaf-labelled trivalent trees are widely used to represent evolutionary relationships in biology; the labels of the leaves correspond to different species. Any binary property of the species (such as the presence or absence of wings) can be represented by a 2-colouring of the leaves of the associated tree T. Such a 2-colouring induces a natural weight called the (parsimony) length of T which is the basis of the maximum parsimony approach for reconstructing trees from discrete data. A natural question in biostatistics asks for the distribution of the parsimony length of a randomly chosen leaf-labelled trivalent tree relative to a random 2-colouration in which each leaf is assigned a colour independently with constant probabilities. In this paper we show that this distribution is asymptotically normal and determine the leading terms of its mean and variance.

1. Introduction

The reconstruction of phylogenetic trees from discrete data is an important problem in evolutionary biology; see, e.g., [19]. Such trees are typically trivalent trees with n $(n \geq 3)$ leaves (endnodes) labelled $1, 2, \ldots, n$ and $n-2$ unlabelled interior

nodes of degree 3. We will write T_n to denote such a tree; there are $(2n-5)!! = 1 \cdot 3 \cdot 5 \cdots (2n-5)$ such trees (up to isomorphism), a result going back at least to [14].

In biology the label set $\{1, 2, \ldots, n\}$ typically corresponds to a set of extant biological species under study, while the interior nodes correspond to hypothetical ancestral species. Discrete biological data generally occur in the form of what biologists refer to as *characters*; these are functions from $\{1, 2, \ldots, n\}$ into some set $\mathcal{C}$ of states, which we will refer to here as colours. A character $\chi : \{1, 2, \ldots, n\} \to \mathcal{C}$ induces a colouration of the leaves of T_n in which the leaf labelled i is assigned the colour $\chi(i)$. Under this interpretation, the *(parsimony) length* of T_n relative to χ is the minimum number $w = w(T_n, \chi)$ of edges of T_n that join nodes with different colours over all the $|\mathcal{C}|^{n-2}$ extensions of χ to a $\mathcal{C}$-colouration of all the nodes of T_n. Note that $0 \leq w(T_n, \chi) \leq n - max\{|\chi^{-1}(c)| : c \in \mathcal{C}\}$, where the upper bound arises by assigning the most frequently occurring colour c to all interior nodes of T_n to obtain a (possibly non-optimal) extension. In particular, $0 \leq w(T_n, \chi) \leq n/2$ when $|\mathcal{C}| = 2$. We remark also that $w(T_n, \chi)$ can be calculated by a $O(n)$ algorithm due to Fitch [6]; see also [7]. An example of a character χ with parsimony length 3 on a tree T_7 is given in Figure 1.

The function w is fundamental to the reconstruction of phylogenetic trees from a sequence of characters by the maximum parsimony approach. Under this approach, a tree is selected that minimizes the sum of $w(T_n, \chi_i)$ over the characters $\chi_1, \chi_2, \ldots, \chi_k$ in the sequence. One rationale for this approach is that it seeks a tree that requires the fewest evolutionary events to explain the observed data on a tree (see [19] for further biological background). In order to assess the statistical significance of results produced by such approaches it is helpful to know the distribution of the function w under simple null models (such an approach was adopted, for example, in [1], [16], and [17]). One simple null model is to select a tree

uniformly at random (from the set of all trivalent trees, leaf labelled from $\{1, 2, \ldots, n\}$) and generate characters randomly by assigning colours to each element of $\{1, 2, \ldots, n\}$ independently and according to the same probability law (which need not necessarily be uniform across $\mathcal{C}$).

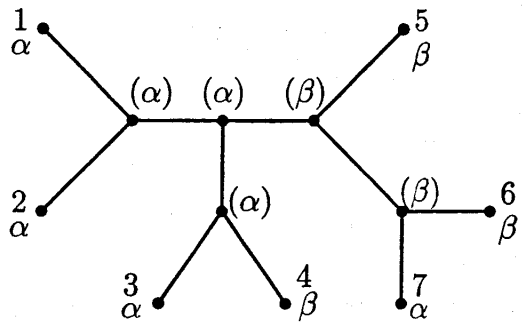

Fig. 1. A trivalent tree T_7, leaf labelled by $\{1, 2, \ldots, 7\}$, together with a character $\chi : \{1, 2, \ldots, 7\} \to \{\alpha, \beta\}$ with $w(T_7, \chi) = 3$. An extension of χ to all the nodes of T_7 is indicated by the values shown in parentheses.

This model was investigated by Hamel and Steel [10] with the aim of calculating the asymptotic mean μn of the distribution of w, where μ is a constant that depends only on $|\mathcal{C}|$ and the probability distribution of colour assignments to each leaf. They gave an explicit formula for μ when $|\mathcal{C}| = 2$ (see equation (1.1) below); and, in general, they showed that μ could be expressed in terms of the (unique) solution of a system of simultaneous quadratic equations. Butler [2] also investigated this model when $|\mathcal{C}| = 2$ to address related but slightly different questions.

In this paper we also consider the case $|\mathcal{C}| = 2$ and show that the distribution of w under the null model described is asymptotically normal, with mean $E(n)$ and $V(n)$ given,

respectively, by

$$E(n) = \mu n + O(1) \quad \text{and} \quad V(n) = \sigma^2 n + O(n^{1/2})$$

where

$$\mu = \tfrac{2}{3}\left\{1 - (1 - 3pq)^{1/2}\right\} \tag{1.1}$$

and

$$\sigma^2 = \mu(1-\mu)(4 - 14\mu + 9\mu^2)(2 - 3\mu)^{-2} \tag{1.2}$$

and where p is the probability that any given leaf is assigned a particular colour from $\mathcal{C}$.

This result complements two earlier central limit theorems involving the parsimony length function w. Moon and Steel [12] established the asymptotic normality of w when $|\mathcal{C}| = 2$ and the tree T_n is selected uniformly at random but where the character is selected so that the proportion of leaves assigned a particular colour from $\mathcal{C}$ is equal to — or very close to — p (we describe this result more formally in the next section). In this case the asymptotic mean and variance are also of the form $\mu' n$ and $s^2 n$, for constants μ' and s, and indeed $\mu' = \mu$; however, $s \neq \sigma$ when $p \neq 1/2$ and, as we shall see, they behave quite differently as functions of p. Steel, Goldstein and Waterman [18] established a different type of central limit theorem involving w. In their paper the tree T_n is fixed (for any given n) instead of being randomly selected; the leaves of the trees are again randomly and independently coloured (but now $\mathcal{C}$ may have more than two elements and the probability distribution on the colours for each leaf may vary from leaf to leaf, subject to a mild technical condition).

These last two central limit theorems can thus be viewed as results in which either the proportion of leaves assigned a particular colour or the tree T_n itself is held constant, whereas

in the present paper we regard both the colourations and the trees as being randomly selected.

2. Preliminaries and Summary

We shall assume henceforth that $\mathcal{C}$ has just two states. Let a and b denote non-negative integers such that $a+b=n$. Let χ'_a denote a 2-colouring that assigns the first colour to a leaves of a leaf labelled trivalent tree T_n and the second colour to the remaining b leaves. For any non-negative integer k, let $p(n,k;a,b)$ denote the probability that $w'=w(T_n,\chi'_a)$ equals k for a tree T_n selected uniformly at random from the set of $(2n-5)!!$ such trees. (We remark that, by symmetry, it follows that the distribution of w' is independent of which a leaves are assigned the first colour.) For example, if $min(a,b)=0$, then $p(n,k;a,b)=1$ if $k=0$ and zero otherwise. In general, if $min(a,b)\neq 0$, then

(2.1)
$$p(n,k;a,b)=2^k\,\frac{(2n-3k)}{(2a-k)(2b-k)}\,\frac{(2a-k)!}{(a-k)!}\times\frac{(2b-k)!}{(b-k)!}\,\frac{(n-k)!}{(k-1)!(2n-2k)!}$$

if $1\leq k\leq min\,(a,b)$ and zero otherwise (see [3], [5], or [15]).

Now let p and q be fixed positive constants such that $p+q=1$. Let us suppose that $|a-pn|=|b-qn|\leq n^{1/3-2\varepsilon}$ for some fixed ε, $0<\varepsilon<1/6$. It was shown in [12] that the distribution of the variable $(w'-\mu' n)/s\sqrt{n}$ is asymptotically normal with zero mean and unit variance, where

(2.2)
$$\mu'=\tfrac{2}{3}\left\{1-(1-3pq)^{1/2}\right\}$$

and

(2.3)
$$s^2=\mu^2(1-\mu)(2-3\mu)^{-2}.$$

Our object here is to consider an extension of this problem. Let χ_p denote a random 2-colouring in which each leaf of T_n is independently assigned the first colour with probability p and the second colour with probability q. Let $P(n,k) = P(n,k;p,q)$ denote the probability that $w = w(T_n, \chi_p)$ equals k over all the $(2n-5)!!$ trivalent trees T_n. In Section 3 we give an expression for $P(n,k)$ as the product of two quantities. In Sections 4 and 5 we consider the asymptotic behaviour of these two quantities; and then in Section 6 we consider the asymptotic behaviour of $P(n,k)$ itself. It will follow from these results that the distribution of the variable $(w-\mu n)/\sigma\sqrt{n}$ is also asymptotically normal with zero mean and unit variance, where μ and σ are as defined in (1.1) and (1.2).

3. An Expression for $P(n,k)$

Before proceeding, we recall that if $y(t) = \frac{1}{2}\{1-(1-4t)^{1/2}\}$, then $y = t(1-y)^{-1}$ and

$$y^h(t) = \sum_{n=h}^{\infty} \frac{h}{2n-h} \binom{2n-h}{n} t^n \tag{3.1}$$

for $h = 1, 2, \ldots$. (See, e.g. [4; p. 164].) In what follows we let $[t^n]Y(t)$ denote the coefficient of t^n in the power series $Y(t)$.

Lemma 1. *Suppose that* $1 \le k \le n/2$ *and let*

$$F_{n,k} = 2^k \frac{2n-3k}{k} \frac{n!}{k!} \frac{(n-k)!}{(2n-2k)!} \tag{3.2}$$

and

$$S_{n,k} = S_{n,k}(p,q) = [t^n]\{\big(y(pt)y(qt)\big)^k\}, \tag{3.3}$$

where

$$y(t) = \tfrac{1}{2}\{1 - (1-4t)^{1/2}\}.$$

Then

$$P(n,k) = F_{n,k} \cdot S_{n,k}. \tag{3.4}$$

Proof. If $a + b = n$, then the probability that the random 2-colouring χ_p assigns the first colour to a leaves of T_n and the second colour to the b remaining leaves is $\binom{n}{q} p^a q^b$. Hence,

$$P(n,k) = \sum_{a+b=n} \binom{n}{a} p^a q^b \cdot p(n,k;a,b).$$

If we substitute the expression for $p(n,k;a,b)$ given in (2.1) into the right hand side of this equation, factor out the quantities that are independent of a and b, and then appeal to (3.1), we find that

$$\begin{aligned} P(n,k) &= 2^k \, \frac{2n-3k}{k} \, \frac{n!}{k!} \, \frac{(n-k)!}{(2n-2k)!} \\ &\qquad \times \sum_{a+b=n} \frac{k}{2a-k} \binom{2a-k}{a} p^a \, \frac{k}{2b-k} \binom{2b-k}{b} q^b \\ &= F_{n,k} \cdot [t^n]\{\big(y(pt)y(qt)\big)^k\} \\ &= F_{n,k} \cdot S_{n,k}, \end{aligned}$$

as required.

Corollary 1.1. *If* $p = q = 1/2$, *then*

$$P(n,k) = \frac{2n-3k}{n-k}\binom{n-k}{k}\left(\frac{1}{2}\right)^{n-k}.$$

Proof. If $p = q = 1/2$, then

$$S_{n,k} = [t^n]\{(y(t/2))^{2k}\} = \frac{k}{n-k}\binom{2n-2k}{n}\left(\frac{1}{2}\right)^n,$$

in view of (3.1). The required conclusion now follows from (3.4).

We remark that Corollary 1.1 is equivalent to Theorem 4 in [15].

4. Estimates for $F_{n,k}$

We observe for later use that it follows readily from definition (1.1) that $\mu \leq 1/3$ with equality holding if and only if $p = q = 1/2$.

Lemma 2. *Let*

$$\mu = \tfrac{2}{3}\{1-(1-3pq)^{1/2}\}$$

and suppose the integers k *and* n *tend to infinity in such a way that*

$$\Delta := k - \mu n = O(n^{2/3})$$

as $k, n \to \infty$. *Then*

(4.1)

$$F_{n,k} = \frac{2-3\mu}{\mu^{3/2}\sqrt{2}}\left(\frac{8(1-\mu)}{\mu}\right)^k\left(\frac{1}{4(1-\mu)}\right)^n \exp\{-\Delta^2/2\mu(1-\mu)n\}$$
$$\times\{1+O(1/n)+O(\Delta/n)+O(\Delta^3/n^2)\}$$

holds uniformly as $k, n \to \infty$. *Furthermore,*

$$(4.2)\quad F_{n,k} = O(1)\Big(\frac{8(1-\mu)}{\mu}\Big)^k \big(\frac{1}{4(1-\mu)}\big)^n \exp\{-\Delta^2/2n\}$$

for all k *and* n, $1 \le k \le n/2$.

Proof. We recall that if r and n are integers tending to infinity in such a way that $r = \rho n + R$, where ρ is a positive constant and $|R/\rho n| < 1/2$, say, then it follows from Stirling's formula and Taylor's theorem that

(4.3)
$$r! = \sqrt{2\pi\rho n}\,(\rho n/e)^r \exp\{R + R^2/2\rho n\}$$
$$\times \{1 + O(1/n) + O(R/n) + O(R^3/n^2)\}$$

as $r, n \to \infty$, where the constants implicit in the O-terms depend only on ρ. When we apply relation (4.3) to the factorials in definition (3.2) and simplify, we obtain conclusion (4.1).

It remains to prove (4.2). Let $N = \lfloor \mu n \rfloor$; it follows from definition (3.2) that if $k = N + j \le n/2$, $j \ge 1$, then

(4.4)
$$F_{n,N+j}/F_{n,N} \le 2^j N^{-j}(2n - 2N)_{2j}/(n-N)_j$$
$$= 8^j N^{-j}(n-N)^j \prod_{i=1}^{j} \{1 - (i - \tfrac{1}{2})/(n-N)\}$$
$$\le 8^j N^{-j}(n-N)^j \exp\Big\{-\sum_{i=1}^{j}(i - \tfrac{1}{2})/n\Big\}$$
$$= O(1)\big(8(1-\mu)/\mu\big)^j \exp\{-j^2/2n\}.$$

And, if $k = N - j, \quad 1 \le j < N$, then

(4.5)

$$F_{n,N-j}/F_{n,N} \le 2^{-j}\,\frac{2n-3N+3j}{2n-3N}\,\frac{N-j+1}{N-j}$$
$$\times\, N^j(n-N+j)_j/(2n-2N+2j)_{2j}$$

$$= O(1)8^{-j}N^j(n-N)^{-j}\prod_{i=1}^{j}\{1+(i-\tfrac{1}{2})/(n-N)\}^{-1}$$

$$\le O(1)8^{-j}N^j(n-N)^{-j}\prod_{i=1}^{j}\{1-(i-\tfrac{1}{2})/n\}$$

$$= O(1)\big(8(1-\mu)/\mu\big)^{-j}\exp\{-j^2/2n\}.$$

When we combine (4.4) and (4.5) and appeal to relation (4.1) with $k = N$ and the fact that

$$(k-N)^2/2n = \Delta^2/2n + O(\Delta/n) = \Delta^2/2n + o(1),$$

we find that

$$F_{n,k} = O(1)\big(8(1-\mu)/\mu\big)^{k-N}\exp\{-(k-N)^2/2n\}F_{n,N}$$

$$= O(1)\big(8(1-\mu)/\mu\big)^{k}\big(4(1-\mu)\big)^{-n}\exp\{-\Delta^2/2n\}.$$

This completes the proof of Lemma 2.

5. An Estimate for $S_{n,k}$

Let $Y(t) = \sum_{m=0}^{\infty} Y_m t^m$ denote a function with non-negative coefficients that is analytic when $|t| < R$, where $0 < R < \infty$. We assume that $Y_m > 0$ for at least three values of m and that $gcd\{m : Y_m > 0\} = 1$. Let

(5.1) $$g(t) := t\,\frac{d}{dt}\,\log Y(t) = tY'(t)/Y(t)$$

and

$$G(t) := tg'(t) = t^2Y''(t)/Y(t) + g(t) - g^2(t) \tag{5.2}$$

for $0 < t < R$. It is not difficult to see that $G(t) > 0$ for $0 < t < R$, so $g(t)$ is strictly increasing in this interval. To obtain an estimate for $S_{n,k}$ we shall make use of a result given in [11] on the behaviour of coefficients in powers of such functions $Y(t)$. The result is, in effect, a version of the classical local limit theorem for sums of suitably defined independent, identically distributed random variables; for related results, see, for example, [8], [9], or some of the other references in [11].

Lemma 3. *Let α be a constant such that $0 < \alpha < g(R^-)$; let η denote the unique number such that $0 < \eta < R$ and*

$$g(\eta) = \alpha. \tag{5.3}$$

Suppose the integers K and N tend to infinity in such a way that

$$D := K - \alpha N = O(N^{2/3})$$

as $K, N \to \infty$. Then

(5.4)

$$[t^K]Y^N(t) = \big(2\pi G(\eta)N\big)^{-1/2}Y^n(\eta)\eta^{-K}\exp\{-D^2/2G(\eta)N\}$$
$$\times\{1 + O(1/N) + O(D/N) + O(D^3/N^2)\}$$

holds uniformly as $K, N \to \infty$.

We now establish certain algebraic relations that we shall use later when we apply Lemma 3 to the problem of estimating $S_{n,k}$.

Lemma 4. *Let*

$$y(t) = \tfrac{1}{2}\{1 - (1-4t)^{1/2}\}$$

and for given $p, q > 0$, *where* $p + q = 1$, *let*

$$\mu := \tfrac{2}{3}\{1 - (1-3pq)^{1/2}\} \tag{5.5}$$

and

$$\eta := \frac{1}{4(1-\mu)} \cdot \tag{5.6}$$

Then

$$y(p\eta) = \frac{1}{2}\left(\frac{p-\mu}{1-\mu}\right)^{1/2}, \tag{5.7}$$

$$y'(p\eta) = \left(\frac{1-\mu}{q-\mu}\right)^{1/2}, \tag{5.8}$$

and

$$y''(p\eta) = 2\left(\frac{1-\mu}{q-\mu}\right)^{3/2}. \tag{5.9}$$

Proof. We noted earlier that $\mu \leq 1/3$. Now

$$1 - 3pq = 1 - 3p + 3p^2 > \left(1 - \tfrac{3}{2}p\right)^2$$

for $p > 0$. So, in particular, if $0 < p \leq 1/3$, then

$$\mu < \frac{2}{3}\{1 - (1 - \tfrac{3}{2}p)\} = p$$

and a similar result holds with respect to q . Consequently,

$$\mu < p \text{ and } \mu < q \tag{5.10}$$

for all $p, q > 0$.

It follows from definition (5.5) that $3\mu^2 - 4\mu + 4pq = 0$. Hence,

$$\mu^2 = 4(\mu^2 - \mu + pq) = 4(p - \mu)(q - \mu),$$

or

$$\mu = 2(p - \mu)^{1/2}(q - \mu)^{1/2}. \tag{5.11}$$

This readily implies that

$$(1 - \mu)^{1/2} = (p - \mu)^{1/2} + (q - \mu)^{1/2}. \tag{5.12}$$

Now, since $\eta = \big(4(1 - \mu)\big)^{-1}$, it follows that

$$1 - 4p\eta = 1 - p/(1 - \mu) = (q - \mu)/(1 - \mu)$$

and, hence,

$$(1 - 4p\eta)^{1/2} = \big((q - \mu)(1 - \mu)\big)^{1/2}. \tag{5.13}$$

Consequently,

$$\begin{aligned} y(p\eta) &= \tfrac{1}{2}\left\{1 - (1 - 4p\eta)^{1/2}\right\} \\ &= \tfrac{1}{2}\left\{1 - \big((q - \mu)/(1 - \mu)\big)^{1/2}\right\} \\ &= \tfrac{1}{2}\big((p - \mu)/(1 - \mu)\big)^{1/2}, \end{aligned} \tag{5.14}$$

by (5.13) and (5.12). This proves (5.7); and (5.8) and (5.9) follow upon substituting relation (5.13) into the expressions for $y'(p\eta)$ and $y''(p\eta)$.

Lemma 5. *Let*

$$\mu = \tfrac{2}{3}\{1-(1-3pq)^{1/2}\} \tag{5.15}$$

and let

$$A = \tfrac{1}{2}\mu^{-3}(1-\mu)(4-14\mu+9\mu^2). \tag{5.16}$$

Suppose the integers k *and* n *tend to infinity in such a way that*

$$\Delta := k-\mu n = O(n^{2/3})$$

as $k,n \to \infty$ *. Then*

(5.17)

$$S_{n,k} = (2\pi\mu An)^{-1/2}\Big(\frac{\mu}{8(1-\mu)}\Big)^k\big(4(1-\mu)\big)^n \, exp\,\{-\Delta^2/2\mu^3 An\}$$

$$\times\{1+O(1/n)+O(\Delta/n)+O(\Delta^3/n^2)\}$$

holds uniformly as $k,n \to \infty$ *.*

Proof. We apply Lemma 3 to the function $Y(t) = y(pt)y(qt)$, where $y(t)$ is as defined earlier, with $N=k$, $K=n$, $\alpha = \mu^{-1}$, and $D = -\mu^{-1}\Delta$. It is easy to see that Y, α, and D and the choice of N and K satisfy the conditions of Lemma 3.

We now show that if $\eta = \big(4(1-\mu)\big)^{-1}$, then $g(\eta) = \mu^{-1}$, so that condition (5.3) holds also. It follows readily from the definitions of $Y(t)$ and $g(t)$ that in the present case

$$g(\eta) = g_p(\eta) + g_q(\eta)$$

where

$$g_p(\eta) = p\eta\, y'(p\eta)/y(p\eta)$$

and similarly for $g_q(\eta)$. When we substitute relations (5.7) and (5.8) into the expression for $g_p(\eta)$ and then appeal to (5.11), we find that

$$g_p(\eta) = \frac{p}{4(1-\mu)} \sqrt{\frac{1-\mu}{q-\mu}}\, 2 \sqrt{\frac{1-\mu}{p-\mu}} = \frac{p}{\mu}\,.$$

Consequently,

$$g(\eta) = g_p(\eta) + g_q(\eta) = (p+q)/\mu = \mu^{-1}, \tag{5.18}$$

as required.

We next observe that

$$\begin{aligned} Y(\eta) &= y(p\eta)y(q\eta) \\ &= \frac{1}{2} \sqrt{\frac{p-\mu}{1-\mu}}\, \frac{1}{2} \sqrt{\frac{q-\mu}{1-\mu}} \\ &= \frac{1}{8}\, \frac{\mu}{1-\mu}\,, \end{aligned} \tag{5.19}$$

in view of relations (5.7) and (5.11).

To evaluate $G(\eta)$ we note that

$$G(\eta) = G_p(\eta) + g_p(\eta) - g_p^2(\eta) + G_q(\eta) + g_q(\eta) - g_q^2(\eta)$$

where

$$G_p(\eta) = (p\eta)^2 y''(p\eta)/y(p\eta)$$

and similarly for $G_q(\eta)$. When we substitute relations (5.9) and (5.7) into the expression for $G_p(\eta)$ and then appeal to (5.11), we find that

(5.20)

$$\begin{aligned} A_p(\eta) &= \Big(\frac{p}{4(1-\mu)}\Big)^2 2\Big(\frac{1-\mu}{q-\mu}\Big)^{3/2} 2\Big(\frac{1-\mu}{p-\mu}\Big)^{1/2} \\ &= 2p^2(p-\mu)\mu^{-3}, \end{aligned}$$

so

$$G_p(\eta) + g_p(\eta) - g_p^2(\eta) = 2p^2(p-\mu)\mu^{-3} + p\mu^{-1} - p^2\mu^{-2}$$
$$= (2p^3 - 3p^2\mu + p\mu^2)\mu^{-3}.$$

A similar relation holds when p is replaced by q. Hence,

(5.21)
$$G(\eta) = \mu^{-3}\{2(p^3+q^3) - 3(p^2+q^2)\mu + (p+q)\mu^2\}$$
$$= \mu^{-3}\{2(1-3pq) - 3(1-2pq)\mu + \mu^2\}$$
$$= \mu^{-3}\{(\mu^2 - 3\mu + 2) - 6pq(1-\mu)\}$$
$$= \mu^{-3}(1-\mu)\{2(1-3pq) - \mu\}$$
$$= \tfrac{1}{2}\mu^{-3}(1-\mu)\{(2-3\mu)^2 - 2\mu\}$$
$$= \tfrac{1}{2}\mu^{-3}(1-\mu)(4 - 14\mu + 9\mu^2) = A,$$

as defined in (5.16), and where we have used definition (5.15) to obtain the next to the last line. It follows, therefore, that

(5.22)
$$NG(\eta) = kA = (\mu n + \Delta)A$$
$$= \mu A n\big(1 + O(\Delta/n)\big)$$

and

(5.23)
$$D^2/2\,G(\eta)N = \Delta^2/2\mu^2 G(\eta)N$$
$$= \Delta^2/2\mu^3 A n\big(1 + O(\Delta(n)\big),$$

by the definitions of N, Δ, and D and (5.21). Conclusion (5.17) now follows from Lemma 3, the definition of η, and relations (5.19), (5.22), and (5.23).

6. Main Results

We are now ready to determine the asymptotic behaviour of $P(n,k)$.

Theorem 1. *Let*

$$\mu = \tfrac{2}{3}\{1-(1-3pq)^{1/2}\} \tag{6.1}$$

and let

$$\sigma^2 = \mu(1-\mu)(4-14\mu+9\mu^2)(2-3\mu)^{-2}. \tag{6.2}$$

Suppose the integers k and n tend to infinity in such a way that

$$\Delta := k-\mu n = O(n^{2/3})$$

as $k, n \to \infty$. Then

(6.3)

$$P(n,k) = \frac{1}{\sigma\sqrt{2\pi n}} \exp\{-\Delta^2/2\sigma^2 n\} \times \{1+O(1/n)+O(\Delta/n)+O(\Delta^3/n^2)\}$$

holds uniformly as $k, n \to \infty$. Furthermore,

$$P(n,k) = O(1)\exp\{-\Delta^2/2n\} \tag{6.4}$$

for all k and n, $1 \le k \le n/2$.

Proof. When we substitute the estimates for $F_{n,k}$ and $S_{n,k}$ given by Lemmas 3 and 5, into expression (3.4) and simplify, bearing in mind the definitions of A and σ^2, we find that

$$P(n,k) = \frac{B}{\sqrt{2\pi n}} \exp\{-C\Delta^2/2n\} \times \{1+O(1/n)+O(\Delta/n)+O(\Delta^3/n^2)\}$$

where

$$B = \frac{2-3\mu}{\sqrt{2\mu^4 A}} = \frac{2-3\mu}{\left(\mu(1-\mu)(4-14\mu+9\mu^2)\right)^{1/2}} = \frac{1}{\sigma}$$

and

$$\begin{aligned} C &= \frac{1}{\mu(1-\mu)} + \frac{1}{\mu^3 A} \\ &= \frac{1}{\mu(1-\mu)} + \frac{2}{(1-\mu)(4-14\mu+9\mu^2)} \\ &= \frac{(2-3\mu)^2}{\mu(1-\mu)(4-14\mu+9\mu^2)} = \frac{1}{\sigma^2}. \end{aligned}$$

This implies conclusion (6.3).

To prove (6.4) when $k > 0$, we note that

$$\begin{aligned} S_{n,k} &= [t^n]Y^k(t) \\ &\le Y^k(\eta)\eta^{-n} \\ &= \left(\frac{\mu}{8(1-\mu)}\right)^k \left(4(1-\mu)\right)^n, \end{aligned} \tag{6.5}$$

appealing to the definition of η and relation (5.19) at the last step. When we combine this with relations (3.4) and (4.2) we obtain conclusion (6.4).

Relation (3.4) does not apply when $k = 0$ so we must consider this case separately. It is easy to see that $P(n,0) = p^n + q^n$. Now $\mu < p < 1$, by (5.10), and $1 - x \le e^{-x}$ for $x \ge 0$. Hence,

$$q = 1 - p \le e^{-p} < e^{-\mu} < e^{-\mu^2}.$$

Similarly,

$$p < e^{-\mu^2},$$

and so

$$P(n,0) < 2e^{-\mu^2 n},$$

which clearly implies (6.4) when $k = 0$. This suffices to complete the proof of the theorem.

We remark that it follows from (6.4) that if

$$H := H(n) = n^{1/2} \log n,$$

then

$$(6.7) \qquad Pr\{|w - \mu n| > H\} = O\big(n \exp\{-H^2/2n\}\big)$$

$$= O(n^{-h})$$

for any positive constant h.

Theorem 2. *Let $E(n)$ and $V(n)$ denote the expected value and the variance of $w = w(T_n, \chi_p)$ over all the $(2n-5)!!$ trivalent trees T_n. Let*

$$\mu = \tfrac{2}{3}\{1 - (1 - 3pq)^{1/2}\}$$

and

$$\sigma^2 = \mu(1-\mu)(4 - 14\mu + 9\mu^2)(2 - 3\mu)^{-2}.$$

Then

$$(6.8) \qquad E(n) = \mu n + O(1) \quad \text{and} \quad V(n) = \sigma^2 n + O(n^{1/2})$$

as $n \to \infty$. Furthermore, if λ is any constant, then

(6.9)

$$Pr\{w(T_n, \chi_p) \le \mu n + \lambda \sigma n^{1/2}\}$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\lambda} e^{-t^2/2} dt + O(n^{-1/2})$$

as $n \to \infty$.

Conclusions (6.8) and (6.9) follow readily from Theorem 1. The contributions from the central part of the distribution may be evaluated by appealing to (6.3) and then using standard error estimates for the difference between a Riemann sum of a function of bounded variation and the corresponding integral. (See, e.g., [13; p. 37, ex. 9] for a closely related result.) The contributions from the tails of the distribution may be disposed of by appealing to (6.7). We omit the details as they are fairly straightforward.

We remark that it follows from the definitions of s^2 and σ^2 that

$$\begin{aligned}\sigma^2/s^2 &= (4 - 14\mu + 9\mu^2)/\mu \\ &= 1 + (1 - 3\mu)(4 - 3\mu)/\mu.\end{aligned}$$

Hence, $\sigma^2 \geq s^2$ with equality holding if and only if $\mu = 1/3$, that is, if and only if $p = q = 1/2$. Furthermore, it can be shown that $\sigma^2 \leq .1296\ldots$ with equality holding if and only if p or q equals $.2505\ldots$.

References

[1] J.W. Archie and J. Felsenstein, *The number of evolutionary steps on random and minimum length trees for random evolutionary data*, Theor. Pop. Biol. **43** (1993), 52-79.

[2] J.P. Butler, *Fraction of trees with given root traits; the limit of large trees*, J. Theor. Biol. **147** (1990), 265-274.

[3] M. Carter, M. Hendy, D. Penny, L.A. Székely, and N.C. Wormald, *On the distribution of lengths of evolutionary trees*, SIAM J. Disc. Math. **3** (1990), 38-47.

[4] L. Comtet, *Advanced Combinatorics*, Reidel, Dordrecht, 1974.

[5] P.L. Erdös and L.S. Székely, *Counting bichromatic evolutionary trees*, Disc. Appl. Math. **47** (1993), 1-8.

[6] W.M. Fitch, *Towards defining the course of evolution: Minimum change for a specific tree topology*, Syst. Zool. **20** (1971), 406-415.

[7] J.A. Hartigan, *Minimum mutation fits to a given tree*, Biometrics **29** (1973), 53-65.

[8] B.V. Gnedenko and A.N. Kolmogorov, *Limit Distributions for Sums of Independent Random Variables*, Addison-Wesley, Cambridge, 1954.

[9] I.J. Good, *Saddle-point methods for the multinomial distribution*, Ann. Math. Stat. **28** (1957), 861-881.

[10] A. Hamel and M.A. Steel, *The length of a leaf colouration on a random binary tree*, SIAM J. Disc. Math. **10** (1997), 359-372.

[11] A. Meir and J. W. Moon, *On the bipartition numbers of random trees, II*, Ars Comb. **51** (1999), 21-31.

[12] J. W. Moon and M.A. Steel, *A limiting theorem for parsimoniously bicoloured trees*, Appl. Math. Lett. **6** (1993), 5-8.

[13] G. Pólya u. G. Szegö, *Aufgaben und Lehrsätzen aus der Analysis, I*, Springer, Berlin, 1970.

[14] E. Schröder, *Vier combinatorishe Probleme*, Zeitschrift für Mathematik und Physik **15** (1870), 361-376.

[15] M. Steel, *Distributions on bicoloured evolutionary trees arising from the principle of parsimony*, Disc. Appl. Math. **41** (1993), 245-261.

[16] M.A. Steel, P.J. Lockhart, and D. Penny, *Confidence in evolutionary trees from biological sequence data*, Nature **364** (1993), 440-442.

[17] M. Steel, M.D. Hendy, and D. Penny, *Significance of the length of the shortest tree*, J. Classification **9** (1992), 71-90.

[18] M. Steel, L. Goldstein and M.S. Waterman, *A central limit theorem for the parsimony score of trees*, Adv. Appl. Prob. **28** (1996), 1051-1071.

[19] D.L. Swofford, G.J. Olsen, P.J. Waddell and D.M. Hillis, *Phylogenetic inference*, Molecular Systematics 2nd ed. (D.M. Hillis, C. Moritz and B.K. Marble, eds.), Sinaur Associates, Sunderland, MA, 1996, pp. 581-607.