Appl. Math. Lett. Vol. 6, No. 4, pp. 5–8, 1993
Printed in Great Britain.

0893-9659/93 $6.00 + 0.00

A LIMITING THEOREM FOR PARSIMONIOUSLY BICOLOURED TREES*

J. W. Moon
Mathematics Department, University of Alberta
Edmonton, Alberta, Canada, T6G 261
M. A. Steel
Mathematics Department, University of Canterbury
Christchurch, New Zealand

(Received and accepted March 1993)

Abstract—The distribution of leaf-bicoloured trivalent trees, according to an induced weight function (a problem which arises in biostatistics), is shown to be asymptotically normal, with explicitly given parameters.

1. INTRODUCTION

Let T_n denote a trivalent tree with n leaves (endnodes) labelled $1, 2, \ldots, n$, and $n-2$ unlabelled interior nodes of degree three; there are $(2n-5)!! = (2n-5)(2n-7)\ldots 3.1$ such trees when $n \geq 3$, a result dating back to 1870 (see [1]). We suppose that the leaves labelled $1, 2, \ldots, a$ are assigned one colour and that the remaining $b = n - a$ leaves are assigned a second colour. If each interior node of T_n is now assigned one of these two colours, then some of the edges of T_n will join nodes of different colour (if $a, b > 0$). The *weight* $w_{a,b} = w_{a,b}(T_n)$ of T_n is the minimum number of such edges, taken over all the 2^{n-2} bicolourings of the interior nodes of T_n. Fitch's algorithm [2] gives an efficient method for calculating $w_{a,b}(T_n)$. The quantity $w_{a,b}(T_n)$ is central to the reconstruction of phylogenetic trees from aligned genetic sequences, and for certain applications (for example [3]) it is useful to be able to calculate the probability $P_{a,b}(k)$ that $w_{a,b}$ equals k, taken over all the $(2n-5)!!$ trivalent trees T_n. It follows from results of Carter *et al.* [4] or Steel [5] that:

$$P_{a,b}(k) = 2^k \cdot \frac{k(2n-3k)}{(2a-k)(2b-k)} \cdot \frac{(2a-k)!}{(a-k)!} \cdot \frac{(2b-k)!}{(b-k)!} \cdot \frac{(n-k)!}{k!(2n-2k)!}, \qquad n = a+b, \tag{1}$$

if $k \leq \min(a, b)$, and zero otherwise. Our object here is to show that the distribution of $w_{a,b}$ is asymptotically normal, subject to certain assumptions. This complements earlier calculations by Butler [6], who derived certain asymptotic probabilities related to $w_{a,b}(T_n)$.

2. THE MAIN RESULT

Theorem. *$P_{a,b}(k)$ is approximated by a normal density with mean μn and variance $s^2 n$ where*

$$\mu := \frac{2}{3}\left\{1 - \left(1 - 3\frac{ab}{n^2}\right)^{1/2}\right\} \tag{2}$$

and

$$s := \frac{\mu(1-\mu)^{1/2}}{2-3\mu}. \tag{3}$$

*We are indebted to Brendon Mckay and Andreas Dress for some helpful comments.

Specifically, let α and $\beta = 1 - \alpha$, denote positive constants such that

$$a = \alpha n + \delta \quad \text{and} \quad b = \beta n - \delta$$

where $a, b \geq 1$ and $|\delta| \leq n^{1/3-2\epsilon}$ for some fixed ϵ, $0 < \epsilon < 1/6$.

Let $x := \frac{k-\mu n}{s\sqrt{n}}$ where $k \leq \min\{a, b\}$. Then provided $|x| \leq n^{1/6-\epsilon}$,

$$P_{a,b}(k) = \frac{1}{s\sqrt{2\pi n}} \cdot e^{-x^2/2}\{1 + O(n^{-3\epsilon})\}, \tag{4}$$

where the constant implicit in the O-term depends only on α.

PROOF. We first observe that

$$\begin{aligned} \frac{k(2n-3k)}{(2a-k)(2b-k)} &= \frac{(\mu + (xs/\sqrt{n}))\,(2 - 3\mu - (3xs/\sqrt{n}))}{(2\alpha - \mu - (xs/\sqrt{n}))\,(2\beta - \mu - (xs/\sqrt{n}))} \\ &= \frac{\mu(2-3\mu)}{(2\alpha-\mu)(2\beta-\mu)} \cdot \{1 + O(n^{-1/3-\epsilon})\}. \end{aligned} \tag{5}$$

(We remark that it follows readily from our assumptions and the definition of μ that all the denominators we encounter will be strictly positive.)

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

$$\begin{aligned} r! &= \sqrt{2\pi r}\left(\frac{r}{e}\right)^r \cdot \{1 + O(r^{-1})\} \\ &= \sqrt{2\pi\rho n}\left(\frac{\rho n}{e}\right)^r \cdot e^{\rho n(1+R/\rho n)\log(1+R/\rho n)} \cdot \left\{1 + O(n^{-1}) + O\left(\frac{R}{n}\right)\right\} \\ &= \sqrt{2\pi\rho n}\left(\frac{\rho n}{e}\right)^r \cdot e^{\rho n(R/\rho n + (1/2)(R/\rho n)^2 + O((R/\rho n)^3))} \cdot \left\{1 + O(n^{-1}) + O\left(\frac{R}{n}\right)\right\} \\ &= \sqrt{2\pi\rho n}\left(\frac{\rho n}{e}\right)^r \cdot e^{R + (1/2)(R^2/\rho n)} \cdot \left\{1 + O(n^{-1}) + O\left(\frac{R}{n}\right) + O\left(\frac{R^2}{n^3}\right)\right\} \end{aligned} \tag{6}$$

as $r, n \to \infty$, where the constants implicit in the O-terms depend only on ρ.

When we apply (6) to the first quotient of factorials in formula (1), and bear in mind the assumptions about δ and $\Delta := k - \mu n$, we find that

$$\begin{aligned} \frac{(2a-k)!}{(a-k)!} &= \left(\frac{n}{e}\right)^a \cdot \frac{(2\alpha-\mu)^{2a-k+(1/2)}}{(\alpha-\mu)^{a-k+(1/2)}} \cdot e^{\delta + \frac{1}{2n}\left\{\frac{(2\delta-\Delta)^2}{2\alpha-\mu} - \frac{(\delta-\Delta)^2}{\alpha-\mu}\right\}} \cdot \{1 + O(n^{-3\epsilon})\} \\ &= \left(\frac{n}{e}\right)^a \cdot \frac{(2\alpha-\mu)^{2a-k+(1/2)}}{(\alpha-\mu)^{a-k+(1/2)}} \cdot e^{\delta - \frac{\alpha\Delta^2}{2n(2\alpha-\mu)(\alpha-\mu)}} \cdot \{1 + O(n^{-3\epsilon})\}. \end{aligned} \tag{7}$$

Similarly,

$$\frac{(2b-k)!}{(b-k)!} = \left(\frac{n}{e}\right)^b \cdot \frac{(2\beta-\mu)^{2b-k+(1/2)}}{(\beta-\mu)^{b-k+(1/2)}} \cdot e^{\delta - \frac{\beta\Delta^2}{2n(2\beta-\mu)(\beta-\mu)}} \cdot \{1 + O(n^{-3\epsilon})\}. \tag{8}$$

And, finally,

$$\frac{(n-k)!}{k!(2n-2k)!} = \frac{1}{\sqrt{4\pi\mu n}} \cdot \left(\frac{e}{n}\right)^n \cdot \frac{e^{-\frac{\Delta^2}{2n\mu(1-\mu)}}}{\mu^k(1-\mu)^{n-k}4^{n-k}} \cdot \{1 + O(n^{-3\epsilon})\}\ . \tag{9}$$

It now follows from (1), (5), (7), (8), and (9), that

$$P(n,k) = \frac{D}{\sqrt{2\pi n}} \cdot A^a \cdot B^b \cdot K^k \cdot e^{-E\Delta^2/2n} \cdot \{1 + O(n^{-3\epsilon})\}, \tag{10}$$

where the constant implicit in the O-term depends only on α, and where

$$A = \frac{(2\alpha-\mu)^2}{4(\alpha-\mu)(1-\mu)}, \quad B = \frac{(2\beta-\mu)^2}{4(\beta-\mu)(1-\mu)}, \quad K = \frac{\alpha-\mu}{2\alpha-\mu}\cdot\frac{\beta-\mu}{2\beta-\mu}\cdot\frac{8(1-\mu)}{\mu},$$

$$D^2 = \frac{\mu(2-3\mu)^2}{2(2\alpha-\mu)(2\beta-\mu)(\alpha-\mu)(\beta-\mu)},$$

and

$$E = \frac{\alpha}{(2\alpha-\mu)(\alpha-\mu)} + \frac{\beta}{(2\beta-\mu)(\beta-\mu)} + \frac{1}{\mu(1-\mu)}.$$

To simplify these expressions, we note that

$$3\mu^2 - 4\mu + 4\alpha\beta = 0, \tag{11}$$

by the definition of μ. Therefore,

$$\begin{aligned}(2\alpha-\mu)^2 &= (4\alpha^2 - 4\alpha\mu + \mu^2) + (3\mu^2 - 4\mu + 4\alpha\beta)\\ &= 4(\mu^2 - (\alpha+1)\mu + \alpha) = 4(\alpha-\mu)(1-\mu),\end{aligned}$$

so $A = 1$ and, similarly, $B = 1$. Furthermore,

$$\begin{aligned}\mu^2 &= \mu^2 + (3\mu^2 - 4\mu + 4\alpha\beta)\\ &= 4(\mu^2 - \mu + \alpha\beta) = 4(\mu-\alpha)(\mu-\beta),\end{aligned} \tag{12}$$

and

$$\begin{aligned}(2\alpha-\mu)(2\beta-\mu) &= (4\alpha\beta - 2\mu + \mu^2) - (3\mu^2 - 4\mu + 4\alpha\beta)\\ &= 2\mu(1-\mu).\end{aligned} \tag{13}$$

Consequently,

$$D^2 = \frac{\mu(2-3\mu)^2}{\mu(1-\mu)\cdot\mu^2} = \frac{1}{s^2}\ .$$

And, finally,

$$\begin{aligned}E &= \frac{\alpha(2\beta^2 - 3\beta\mu + \mu^2) + \beta(2\alpha^2 - 3\alpha\mu + \mu^2)}{(2\alpha-\mu)(2\beta-\mu)(\alpha-\mu)(\beta-\mu)} + \frac{1}{\mu(1-\mu)}\\ &= \frac{2(\mu^2 - 6\alpha\beta\mu + 2\alpha\beta)}{\mu^3(1-\mu)} + \frac{1}{\mu(1-\mu)} = \frac{3\mu^2 + 4\alpha\beta - 12\alpha\beta\mu}{\mu^3(1-\mu)}\\ &= \frac{4\mu(1-3\alpha\beta)}{\mu^3(1-\mu)} = \frac{(2-3\mu)^2}{\mu^2(1-\mu)} = \frac{1}{s^2},\end{aligned}$$

where we have used (12) and (13) again in the second line, and (11) and the definition of μ in the last two lines. When we replace A, B, K, D, and $E\Delta^2/n$ in relation (10) by $1, 1, 1, s^{-1}$ and x^2, respectively, we obtain the required result.

COROLLARY.

(1) *If z is any constant, then*

$$\Pr\left\{w_{a,b}(T_n) \le un + zsn^{(1/2)}\right\} \to \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-(1/2)x^2}dx \qquad \text{as } n \to \infty.$$

(2) *If $E(n)$ and $V(n)$ denote the mean and variance of $w_{a,b}(T_n)$, then*

$$\frac{E(n)}{n} \to \mu \quad \text{and} \quad \frac{V(n)}{n} \to s^2 \qquad \text{as } n \to \infty.$$

These results follow from (4) by standard arguments that involve approximating appropriate sums by integrals [7, pp. 149–157]. The contributions from values of k such that $|k - \mu n| \ge 2s\sqrt{n\log n}$, say, are negligible; this follows from (4) and the fact—a consequence of (1)—that the probabilities $P_{a,b}(k)$ decrease as $|k - \mu n| = |\Delta|$ increases, at least when $|\Delta| \ge \max\{|\delta|, 1\}$.

3. REMARKS

REMARK 1. An *edge-rooted* trivalent tree is a trivalent tree T_n with a subdivided edge (the new node being the root). Let $B_1(n)$ denote the proportion of the $(2n-3)!!$ edge-rooted trivalent trees T_n for which the root receives the first colour under all minimal-weight bicolourings of the interior nodes of T_n (the leaves of T_n being bicoloured as in the Introduction). Butler [6] investigated the limiting behaviour of $B_1(n)$ and showed, given certain tacit assumptions, that

$$B_1(n) \to \frac{2\alpha - 3\mu}{2 - 3\mu}, \qquad \text{as } n \to \infty.$$

We note here that this result can be easily derived by applying our theorem (and the comment following the corollary) to the identity [8, Theorem 3, equation (5)]

$$B_1(n) = \sum_k \frac{(2a-2k)}{(2n-3k)} P_{a,b}(k),$$

since $\frac{2a-2k}{2n-3k}$ is necessarily bounded above by one for all k, and is uniformly convergent to $\frac{2\alpha-2\mu}{2-3\mu}$ when $|k - \mu n| \leq 2s\sqrt{n \log n}$, say.

REMARK 2. We conjecture that the asymptotic normality described above extends from bicolourings to r-colourings, for $r > 2$, although μ and s may no longer be explicitly-representable functions. A related conjecture, due to M. Waterman and L. Goldstein (personal communication) asserts that for a *fixed* T_n with leaves regarded as i.i.d. random variables which take values in a set of r colours, then the weight of this random leaf colouration of T is asymptotically normal (as $n \to \infty$). We remark here that our theorem allows a proof of this conjecture in the special case when $r = 2$, and the probability of assigning each colour to a leaf is 0.5. This relies on the fact that the number of ways to colour the leaves of a tree T_n with two colours such that the resulting leaf-colouration has weight k on T_n depends only on n and k (see [5]).

REFERENCES

1. E. Schröder, Vier combinatorishe Probleme, *Zeitschrift für Mathematik und Physik* **15**, 361–376 (1870).
2. J.A. Hartigan, Minimum mutation fits to a given tree, *Biometrics* **29**, 53–65 (1973).
3. M. Steel, M.D. Hendy and D. Penny, Significance of the length of the shortest tree, *J. Classification* **9**, 71–90 (1992).
4. 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** (1), 38–47 (1990).
5. M. Steel, Distributions on bicoloured evolutionary trees arising from the principle of parsimony, *Disc. Appl. Math.* (1992) (to appear).
6. J.P. Butler, Fraction of trees with given root traits; the limit of large trees, *J. Theor. Biol.* **147**, 265–274 (1990).
7. A. Rényi, *Probability Theory*, North-Holland, Amsterdam, (1970).
8. M. Steel, Decompositions of leaf-coloured binary trees, *Adv. Appl. Math.* (1992) (to appear).