The Annals of Applied Probability
1991, Vol. 1, No. 3, 406–418

ON THE DISTRIBUTION OF LEAVES IN ROOTED SUBTREES OF RECURSIVE TREES

BY HOSAM M. MAHMOUD AND R. T. SMYTHE

George Washington University

We study the structure of $T_n^{(k)}$, the subtree rooted at k in a random recursive tree of order n, under the assumption that k is fixed and $n \to \infty$. Employing generalized Pólya urn models, exact and limiting distributions are derived for the size, the number of leaves and the number of internal nodes of $T_n^{(k)}$. The exact distributions are given by intricate formulas involving Eulerian numbers, but a recursive argument based on the urn model suffices for establishing the first two moments of the above-mentioned random variables. Known results show that the limiting distribution of the size of $T_n^{(k)}$, normalized by dividing by n is Beta$(1, k-1)$. A martingale central limit argument is used to show that the difference between the number of leaves and the number of internal nodes of $T_n^{(k)}$, suitably normalized, converges to a mixture of normals with a Beta$(1, k-1)$ as the mixing density. The last result allows an easy determination of limiting distributions of suitably normalized versions of the number of leaves and the number of internal nodes of $T_n^{(k)}$.

1. Introduction. A tree on n vertices labeled $1, 2, \ldots, n$ is a rooted recursive tree of order n if the node labeled 1 is distinguished as the root, and for each k, $2 \le k \le n$, the labels of the vertices in the unique path joining the root to the vertex labeled k form an increasing sequence. Within one level the children are conventionally ordered from left to right in an increasing sequence. Figure 1 illustrates all recursive trees of order 4.

The usual model of randomness on the space of recursive trees of order n is the uniform one; that is, we assume all $(n-1)!$ recursive trees to be equally likely [Moon (1974)]. Another way to view this probability space arises from the algorithmic development: Assume a recursive tree of order n evolves from the recursive tree T_{n-1}, of order $n-1$, by choosing a node (a parent) of T_{n-1} at random and joining a node labeled n (a child) to it, all $n-1$ nodes of T_{n-1} being equally likely. The *leaves* of the tree are the childless nodes. This structure arises naturally in the pyramidal hierarchy of companies where a person founds the company, then goes out to recruit new employees [Gastwirth (1977)]. Each entrant in turn competes with the existing work force of the company in recruiting the next employee, the incentive being the promise of a profit proportional to the number of recruits. Moon (1974) suggested recursive trees as a model for the spread of an epidemic. Najock and

Received May 1990; revised September 1990.
AMS 1980 *subject classifications*. Primary 05C05; secondary 60G42, 68E05.
Key words and phrases. Recursive trees, rooted subtrees, generalized Pólya urn models, martingale central limit theorem.

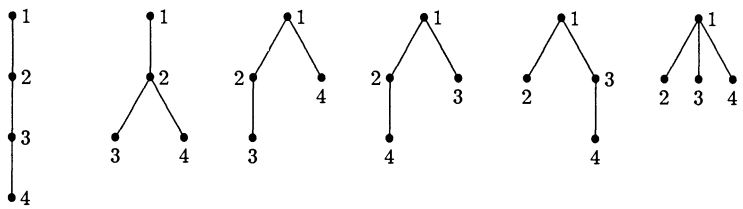

FIG. 1. *The recursive trees of order* 4.

Heyde (1982) used recursive trees to model the family trees of preserved copies of ancient or medieval texts.

The number of leaves in a recursive tree is a random variable of obvious interest. In pyramid schemes, for example, it represents the number of "shut-outs" (those who join the pyramid but fail to recruit anyone). Na and Rapoport (1970) investigated the average number of leaves; Najock and Heyde (1982) found the exact distribution.

In this investigation, we concentrate on the subtree of T_n rooted at the kth node, denoted by $T_n^{(k)}$, for $k = 1, 2, \ldots, n$. (The results of Najock and Heyde then correspond to the special case $k = 1$.) In a pyramid scheme, the founder of a subtree may be (like the original founder) at risk from legal action by the shut-outs of his own subtree. In the problem considered by Najock and Heyde (1982), the leaves of $T_n^{(k)}$ correspond to the terminal copies of a given (nonoriginal) copy. The distribution of the leaves of a rooted subtree is thus of interest. Gastwirth and Bhattacharya (1984) derived the exact distribution of the size of $T_n^{(k)}$ and considered the distribution when k and n go to ∞ in a constant ratio. Here, we fix k and let n go to ∞. The exact distribution of the number of leaves is derived, along with its first two moments, and a central limit theorem is proved for the difference between the number of nonleaf nodes (the internal nodes) in $T_n^{(k)}$ and the number of leaves in that subtree. The last result allows easy determination of the asymptotic distribution of the number of leaves in the subtree rooted at k.

2. Generalized Pólya urn models in recursive trees. A generalized Pólya urn is an urn containing k types of balls with colors $C_1, \ldots, C_k$. Initially, the urn contains a known number of each color. A ball is drawn at random (all balls in the urn being equally likely), the color of the ball is observed, and if its color is C_i then the ball is returned to the urn with α_{ij} additional balls of color C_j, for $j = 1, \ldots, k$. The process is then repeated n times. In the most general scheme the numbers α_{ij} are random variables; for our purposes, however, they will be deterministic. The issue of interest is the composition of the urn after n drawings.

To monitor the growth of $T_n^{(k)}$, we need two colors (black, B and red, R) to distinguish between leaves and internal nodes of $T_n^{(k)}$, respectively; and a third color (white, W) to further distinguish the nodes that are not in that subtree.

All the nodes up to $k-1$ are white, and when k joins the tree a black node is added. We think of this as the initial composition of a generalized Pólya urn having $k-1$ white balls, 1 black ball and no red balls. Let W_n, B_n, R_n denote the number of white, black and red balls after n draws, respectively; the initial urn composition is $W_0 = k-1$, $B_0 = 1$, $R_0 = 0$. We count draws from this point on, so that after n draws the entire tree has $n+k$ nodes, with $B_n + R_n$ nodes in the subtree and W_n nodes outside the subtree ($W_n + B_n + R_n = n + k$).

After the nth stage, the algorithmic development of the tree is as follows: If the parent node picked (ball drawn) is not in the subtree (white), a node is added outside the subtree (white ball added to the urn), and the subtree is unchanged. If the parent node picked (ball drawn) is an internal node of the subtree (red), it remains an internal node in the subtree and a new leaf (black ball) is added. On the other hand, if the parent node picked (ball drawn) is black, it is converted to an internal node (red) of the subtree and a leaf (black) is added to the subtree. The net effect of the last transaction is to add a new internal node (red ball), while the number of leaves (black balls) stays constant. The addition of balls thus follows the scheme:

(2.1)

		Ball picked		
		W	B	R
Balls added	W	1	0	0
	B	0	0	1
	R	0	1	0

and the object of interest is B_n, the number of leaves of the subtree (black balls in the urn) after n draws.

Note that if $k = 1$, all nodes are in the subtree, so there are only two colors and the urn model specializes to Friedman's urn [Friedman (1949)]. Asymptotic results for this urn were developed by Freedman (1965) and rediscovered, by a different route, by Najock and Heyde (1982).

3. The exact distribution of B_n. Let the number of nodes in the subtree after stage n (corresponding to n draws from the urn) be denoted by C_n. Thus

$$C_n = R_n + B_n, \qquad C_0 = 1.$$

We condition on C_n. Following Najock and Heyde,

(3.1)
$$\begin{aligned} P(B_{n+1} = j, C_n = m) &= P(B_{n+1} = j, C_n = m, B_n = j) \\ &\quad + P(B_{n+1} = j, C_n = m, B_n = j-1) \\ &= P(B_{n+1} = j \mid C_n = m, B_n = j) \\ &\quad \times P(B_n = j \mid C_n = m) P(C_n = m) \\ &\quad + P(B_{n+1} = j \mid C_n = m, B_n = j-1) \\ &\quad \times P(B_n = j-1 \mid C_n = m) P(C_n = m). \end{aligned}$$

If $B_n = j$ and $C_n = m$, the numbers of leaves of $T_n^{(k)}$ will not change either if one of the j leaves of $T_n^{(k)}$ is selected as parent or if one of the W_n nodes outside the subtree is selected as parent. Since $W_n + C_n = n + k$, the probability of this is $(n + k + j - m)/(n + k)$. If $B_n = j - 1$, and $C_n = m$, the number of leaves of $T_n^{(k)}$ will increase by one if one of the internal nodes of the subtree is selected as parent; the probability of this is $(m - (j - 1))/(n + k)$. Hence (3.1) gives

$$P(B_{n+1} = j, C_n = m) = \frac{P(C_n = m)}{n + k}\{(n + k + j - m)P(B_n = j \mid C_n = m) + (m - j + 1)P(B_n = j - 1 \mid C_n = m)\}.$$

Now at each stage, any node of the subtree has the same probability of being selected as parent; so the probabilities $P(B_n = j \mid C_n = m)$ and $P(B_n = j - 1 \mid C_n = m)$ are just the probabilities found by Najock and Heyde for a tree with m nodes. Thus

$$P(B_{n+1} = j) = \sum_{m=j}^{n+1} \frac{P(C_n = m)}{n + k} \times \left\{(n + k + j - m)\frac{\left\langle {m - 1 \atop j} \right\rangle}{(m - 1)!} + (m - j + 1)\frac{\left\langle {m - 1 \atop j - 1} \right\rangle}{(m - 1)!}\right\}, \tag{3.2}$$

where $\left\langle {s - 1 \atop t} \right\rangle$, for integers $s \geq 1$ and $t \geq 0$, is the Eulerian number $\sum_{j=0}^{t}(-1)^j(t - j)^{s-1}\binom{s}{j}$ [Knuth (1973), page 37]. But the probabilities $P(C_n = m)$ are known, for by merging red and black together as one type the rule (2.1) clearly shows that C_n is just the number of balls of one type in a standard Pólya urn with $C_0 = 1$, $W_0 = k - 1$. Hence [Johnson and Kotz (1977), page 177]

$$P(C_n = m) = \frac{n!(k - 1)}{(n - m + 1)!(n + k - m)\cdots(n + k - 1)}, \qquad k \geq 2,$$

$$P(C_n = n + 1) = 1, \qquad k = 1.$$

Applying the recurrence [Knuth (1973), page 35]

$$\left\langle {s \atop j} \right\rangle = j\left\langle {s - 1 \atop j} \right\rangle + (s - j + 1)\left\langle {s - 1 \atop j - 1} \right\rangle, \tag{3.3}$$

the relation (3.2) becomes, with a little rearranging,

$$P(B_{n+1} = j) = \frac{(k - 1)}{(n + k)!}\sum_{m=j}^{n+1}(n + k - m - 1)!\binom{n}{m - 1} \times \left\{\left\langle {m \atop j} \right\rangle + (n + k - m)\left\langle {m - 1 \atop j} \right\rangle\right\}. \tag{3.4}$$

To relate this back to the original tree, let

$L_n^{(k)}$ = number of leaves in the subtree rooted at k when the tree size is n

= number of black balls in the urn when the total number of balls is n

= number of black balls after $n - k$ draws from the urn,

since we started with k balls. Hence

$$L_n^{(k)} = B_{n-k}, \tag{3.5}$$

and from (3.4),

$$P\left(L_n^{(k)} = j\right) = \frac{(k-1)}{(n-1)!} \sum_{m=j}^{n-k} (n-m-2)! \binom{n-k-1}{m-1} \times \left\{ \left\langle \begin{matrix} m \\ j \end{matrix} \right\rangle + (n-m-1) \left\langle \begin{matrix} m-1 \\ j \end{matrix} \right\rangle \right\},$$

for $2 \le k \le n$. For $k = 1$, we recover the result of Najock and Heyde by an application of the recurrence (3.3) in (3.2).

4. Moments of B_n. The formula (3.4) does not lend itself readily to closed-form calculation of the moments of B_n. However, recursive arguments combined with known results for the Pólya urn will provide the results fairly easily.

Define the auxiliary random variables ρ_n and β_n as follows:

$$\rho_n = \begin{cases} 1 & \text{if a red ball is drawn on the } n\text{th draw,} \\ 0 & \text{otherwise,} \end{cases}$$

$$\beta_n = \begin{cases} 1 & \text{if a black ball is drawn on the } n\text{th draw,} \\ 0 & \text{otherwise.} \end{cases}$$

According to the algorithmic scheme, a black ball is added to the urn if a red ball is drawn, and vice versa, so that

$$\begin{aligned} B_{n+1} &= B_n + \rho_{n+1}, \\ R_{n+1} &= R_n + \beta_{n+1}, \end{aligned} \tag{4.1}$$

and consequently,

$$\begin{aligned} \mathbf{E}(B_{n+1}) &= \mathbf{E}(B_n) + \mathbf{E}(\rho_{n+1}), \\ \mathbf{E}(R_{n+1}) &= \mathbf{E}(R_n) + \mathbf{E}(\beta_{n+1}). \end{aligned}$$

But

$$\mathbf{E}(\rho_{n+1} \mid W_n, B_n, R_n) = \frac{R_n}{W_n + B_n + R_n} = \frac{R_n}{n+k},$$

so

$$\mathbf{E}(\rho_{n+1}) = \frac{\mathbf{E}(R_n)}{n+k}. \tag{4.2}$$

Similarly,

$$\mathbf{E}(\beta_{n+1}) = \frac{\mathbf{E}(B_n)}{n+k}, \tag{4.3}$$

and we have the simultaneous recurrences

$$\begin{aligned} \mathbf{E}(B_{n+1}) &= \mathbf{E}(B_n) + \frac{\mathbf{E}(R_n)}{n+k}, \\ \mathbf{E}(R_{n+1}) &= \mathbf{E}(R_n) + \frac{\mathbf{E}(B_n)}{n+k}. \end{aligned} \tag{4.4}$$

Recalling that $C_n \equiv R_n + B_n$, with $C_0 = 1$, it is known [Johnson and Kotz (1977), page 179] that

$$\mathbf{E}(C_n) = 1 + \frac{n}{k}$$

or

$$\mathbf{E}(B_n) + \mathbf{E}(R_n) = \frac{n+k}{k}. \tag{4.5}$$

Substitution in (4.4) gives a recurrence relation for $\mathbf{E}(R_n)$ with solution

$$\mathbf{E}(R_n) = \frac{(n+k-1)(n+k) - k(k-1)}{2k(n+k-1)}.$$

From (4.5),

$$\mathbf{E}(B_n) = \frac{(n+k-1)(n+k) + k(k-1)}{2k(n+k-1)}.$$

For large n, it follows that, with k fixed

$$\begin{aligned} \mathbf{E}(B_n) &\sim \frac{n}{2k}, \\ \mathbf{E}(R_n) &\sim \frac{n}{2k}, \end{aligned}$$

and the total size of the subtree is about n/k. From the correspondences (3.5),

$$\mathbf{E}\left(L_n^{(k)}\right) = \frac{n}{2k} + \frac{k-1}{2(n-1)},$$

a result proved independently by Szymański (1990).

For the second moments, we again begin with (4.1), and get

$$\begin{aligned} B_{n+1}^2 &= B_n^2 + \rho_{n+1} + 2B_n\rho_{n+1}, \\ R_{n+1}^2 &= R_n^2 + \beta_{n+1} + 2R_n\beta_{n+1}, \end{aligned}$$

and using (4.2) and (4.3),

$$\begin{aligned}\mathbf{E}(B_{n+1}^2) &= \mathbf{E}(B_n^2) + \frac{\mathbf{E}(R_n)}{n+k} + 2\mathbf{E}(B_n\rho_{n+1}),\\ \mathbf{E}(R_{n+1}^2) &= \mathbf{E}(R_n^2) + \frac{\mathbf{E}(R_n)}{n+k} + 2\mathbf{E}(R_n\beta_{n+1}).\end{aligned} \tag{4.6}$$

Now

$$\mathbf{E}(B_n\rho_{n+1}) = \mathbf{E}(B_n\mathbf{E}(\rho_{n+1}\mid W_n, B_n, R_n)) = \mathbf{E}\left(\frac{B_nR_n}{n+k}\right),$$

from the argument immediately preceding (4.2), and similarly

$$\mathbf{E}(R_n\beta_{n+1}) = \mathbf{E}(R_n\mathbf{E}(\beta_{n+1}\mid W_n, B_n, R_n)) = \mathbf{E}\left(\frac{R_nB_n}{n+k}\right).$$

Hence (4.6) becomes

$$\begin{aligned}\mathbf{E}(B_{n+1}^2) &= \mathbf{E}(B_n^2) + \frac{\mathbf{E}(R_n)}{n+k} + \frac{2\mathbf{E}(B_nR_n)}{n+k},\\ \mathbf{E}(R_{n+1}^2) &= \mathbf{E}(R_n^2) + \frac{\mathbf{E}(B_n)}{n+k} + \frac{2\mathbf{E}(B_nR_n)}{n+k}.\end{aligned} \tag{4.7}$$

Let $V_n \equiv \mathbf{E}(B_n^2 - R_n^2)$, $U_n \equiv \mathbf{E}(B_n^2 + R_n^2)$; clearly $V_0 = U_0 = 1$. From (4.7) and (4.5),

$$V_{n+1} = V_n + \frac{1}{n+k}(\mathbf{E}(R_n) - \mathbf{E}(B_n)) = V_n - \frac{(k-1)}{(n+k)(n+k-1)},$$

which has the solution

$$V_n = \frac{k-1}{n+k-1}, \qquad k \geq 1, n \geq 0. \tag{4.8}$$

For U_n, (4.7) and (4.5) give

$$U_{n+1} = U_n + \frac{1}{k} + 4\frac{\mathbf{E}(R_nB_n)}{n+k} \tag{4.9}$$

and

$$\mathbf{E}(R_nB_n) = \frac{\mathbf{E}(C_n^2) - U_n}{2}. \tag{4.10}$$

From Johnson and Kotz (1977), page 179, we have

$$\mathbf{E}(C_n^2) = \frac{3n}{k} + \frac{2n(n-1)}{k(k+1)} + 1,$$

so (4.9) and (4.10) give

$$U_{n+1} = \left(\frac{n+k-2}{n+k}\right)U_n + \left(\frac{3}{k} + \frac{4n}{k(k+1)}\right), \qquad n = 0, 1, \ldots,$$

which has as solution

$$U_n = \frac{(n+k)(3n+2k+2)}{3k(k+1)} + \frac{(k-2)(k-1)}{3(n+k-2)(n+k-1)} \quad \text{for } n \geq 0. \tag{4.11}$$

From (4.8) and (4.12),

$$\begin{aligned}\mathbf{E}(B_n^2) &= \frac{1}{2}(U_n + V_n) \\ &= \frac{1}{2}\left\{\frac{k-1}{n+k-1} + \frac{(n+k)(3n+2k+2)}{3k(k+1)} + \frac{(k-2)(k-1)}{3(n+k-2)(n+k-1)}\right\},\end{aligned}$$

$$\begin{aligned}\mathbf{E}(R_n^2) &= \frac{1}{2}(U_n - V_n) \\ &= \frac{1}{2}\left\{-\frac{k-1}{n+k-1} + \frac{(n+k)(3n+2k+2)}{3k(k+1)} + \frac{(k-2)(k-1)}{3(n+k-2)(n+k-1)}\right\}.\end{aligned}$$

Using (3.5) again,

$$\mathbf{E}(L_n^{(k)})^2 = \frac{1}{2}\left\{\frac{k-1}{n-1} + \frac{n(3n-k+2)}{3k(k+1)} + \frac{(k-2)(k-1)}{3(n-2)(n-1)}\right\}.$$

For large n and fixed $k > 1$,

$$\mathbf{E}(L_n^{(k)})^2 \sim \frac{n^2}{2k(k+1)}, \qquad \operatorname{Var}(L_n^{(k)}) \sim \frac{n^2(k-1)}{4k^2(k+1)}.$$

For $k = 1$, $\operatorname{Var}(L_n^{(k)}/n) \to 0$, since $L_n^{(k)}/n \to_P 1/2$ in this case [Freedman (1965)].

5. Asymptotic behavior of B_n and R_n. Our starting point is a limit theorem for Pólya urn schemes [Athreya and Ney (1972), page 220]:

THEOREM 5.1. *There is a random variable Y with a* Beta$(1, k-1)$ *distribution such that*

$$\lim_{n\to\infty}\left(\frac{R_n + B_n}{n+k}\right) = Y, \quad a.s.$$

(Of course, if $k = 1$, the limit of the left-hand side is just the constant 1.)

In this section we investigate the asymptotic behavior of $X_n \equiv B_n - R_n$, when k is fixed. Our main result is the following theorem.

THEOREM 5.2. *Let Y be a* Beta$(1, k-1)$ *random variable. Then for $k > 1$ fixed, $\sqrt{3/n}\, X_n \to_{\mathscr{D}} Z$ as $n \uparrow \infty$, where the characteristic function of Z is*

$$\mathbf{E}(e^{-Yt^2/2}) = (k-1)\int_0^1 e^{-yt^2/2}(1-y)^{k-2}\,dy.$$

The limiting distribution of X_n is thus not normal, but a mixture of normals, with Beta$(1, k-1)$ *as the mixing density.*

PROOF. Let $\Delta X_n \equiv X_n - X_{n-1}$, $n \geq 1$ (with $X_0 = 1$). Let $\mathscr{F}_n \equiv \sigma(B_i, R_i, W_i\colon 0 \leq i \leq n)$; then $\mathscr{F}_n \subseteq \mathscr{F}_{n+1}$ for each n and after the nth draw from the urn,

$$\Delta X_n = \begin{cases} +1 & \text{if a red ball is drawn,} \\ 0 & \text{if a white ball is drawn,} \\ -1 & \text{if a black ball is drawn.} \end{cases}$$

Thus

$$\mathbf{E}(\Delta X_n \mid \mathscr{F}_{n-1}) = \frac{R_{n-1} - B_{n-1}}{n+k-1} = -\frac{X_{n-1}}{n+k-1}, \qquad n > 1,$$

$$\mathbf{E}(\Delta X_1 \mid \mathscr{F}_0) = \mathbf{E}(\Delta X_1) = -\frac{1}{k},$$

and

$$\Delta M_n \equiv (X_n - X_{n-1}) + \frac{X_{n-1}}{n+k-1},$$

is a martingale difference sequence. Let $b_{in} \equiv (i+k-1)/(n+k-1)$, $1 \leq i \leq n$, and note that

$$\frac{\max b_{in}^2}{\sum_{i=1}^n b_{in}^2} \to_{n\uparrow\infty} 0. \tag{5.1}$$

An easy calculation shows that

$$\sum_{i=1}^n b_{in}\,\Delta M_i = X_n - \frac{k-1}{n+k-1}. \tag{5.2}$$

Let $s_n^2 \equiv \sum_{i=1}^n b_{in}^2$. Then for each n, $z_{in} \equiv b_{in}\,\Delta M_i/s_n$, $1 \leq i \leq n$, is a martingale difference array. We have

$$z_{in}^2 = \frac{b_{in}^2}{s_n^2}\left\{(X_i - X_{i-1})^2 + \left(\frac{X_{i-1}}{i+k-1}\right)^2 + \frac{2X_{i-1}}{i+k-1}(X_i - X_{i-1})\right\}, \qquad i > 1. \tag{5.3}$$

Now

$$\mathbf{E}(X_i \mid \mathscr{F}_{i-1}) = \mathbf{E}(\Delta M_i \mid \mathscr{F}_{i-1}) + X_{i-1} - \frac{X_{i-1}}{i+k-1} = \frac{i+k-2}{i+k-1} X_{i-1},$$

so that

$$\mathbf{E}(X_{i-1}X_i \mid \mathscr{F}_{i-1}) = \frac{i+k-2}{i+k-1} X_{i-1}^2. \tag{5.4}$$

Hence

$$\begin{aligned}\mathbf{E}(z_{in}^2 \mid \mathscr{F}_{i-1}) &= \frac{b_{in}^2}{s_n^2}\left\{\mathbf{E}\big((X_i - X_{i-1})^2 \mid \mathscr{F}_{i-1}\big) + \frac{X_{i-1}^2}{(i+k-1)^2}\right. \\ &\qquad \left. - \frac{2X_{i-1}^2}{i+k-1} + \frac{2(i+k-2)}{(i+k-1)^2} X_{i-1}^2\right\} \qquad (5.5) \\ &= \frac{b_{in}^2}{s_n^2}\left\{\frac{R_{i-1} + B_{i-1}}{i+k-1} - \frac{X_{i-1}^2}{(i+k-1)^2}\right\}\end{aligned}$$

and

$$\mathbf{E}(z_{in}^2) = \frac{b_{in}^2}{s_n^2}\left\{\frac{\mathbf{E}(R_{i-1}) + \mathbf{E}(B_{i-1})}{i+k-1} - \frac{\mathbf{E}(X_{i-1}^2)}{(i+k-1)^2}\right\}.$$

Equating $\mathbf{E}(X_i - X_{i-1})^2$ to $(\mathbf{E}(R_{i-1}) + \mathbf{E}(B_{i-1}))/(i+k-1)$ and using (4.5) and (5.4), we get the recursion

$$\mathbf{E}(X_i^2) = \left(\frac{i+k-3}{i+k-1}\right)\mathbf{E}(X_{i-1}^2) + \frac{1}{k}.$$

It follows easily that $\mathbf{E}(X_i^2) \sim i/(3k)$ as $i \to \infty$, so that

$$\frac{X_{i-1}^2}{(i+k-1)^2} \to_P 0. \tag{5.6}$$

From (5.1), (5.5), (5.6) and Theorem 5.1, and a standard application of Toeplitz' lemma,

$$\sum_{i=1}^{n} \mathbf{E}(z_{in}^2 \mid \mathscr{F}_{i-1}) \to_P Y,$$

where Y is a Beta$(1, k-1)$ random variable. By (5.1) and (5.6), the conditional Lindeberg condition clearly holds, and we may apply Corollary (3.1) of Hall and Heyde (1980) to the martingale difference array $\{z_{in}\}$ to conclude that

$$\sum_{i=1}^{n} z_{in} \to_{\mathscr{D}} Z,$$

where the characteristic function of Z is $\mathbf{E}(\exp(-Yt^2/2))$. From (5.2) it follows

that

$$\sum_{i=1}^{n} z_{in} = \frac{1}{s_n}\left\{X_n - \frac{k-1}{n+k-1}\right\},$$

and $s_n \sim \sqrt{n/3}$, completing the proof. □

From Theorem 5.2 it follows easily that

$$\frac{R_n - B_n}{n+k} \to_P 0,$$

so that from Theorems 5.1 and 5.2 we have the following corollary.

COROLLARY 5.1.

$$\frac{2R_n}{n} \to_P Y,$$

$$\frac{2B_n}{n} \to_P Y,$$

where Y is a Beta$(1, k-1)$ *random variable. The number of leaves in the subtree rooted at k thus satisfies*

$$\frac{2L_n^{(k)}}{n} \to_P Y.$$

The number of internal nodes of the subtree satisfies a similar limit law.

Note: We are indebted to a referee for pointing out that the result of Corollary 5.1 may be derived from Theorem 5.1 without the use of Theorem 5.2. This follows from an observation made in Section 3—that conditional on $|T_n^{(k)}| = m$, the leaves of the subtree $T_n^{(k)}$ are distributed as the leaves in $T_m^{(1)}$ —and the result of Friedman (1965) that the number of leaves in $T_m^{(1)}$ is asymptotic to $m/2$.

Figure 2 illustrates the exact distribution of $2L_{40}^{(2)}/40$ in comparison with the limiting Beta(1, 1) distribution of $2L_n^{(2)}/n$ as $n \to \infty$. The exact distribution is depicted as a histogram of $n - k = 38$ boxes, each of width $2/n = 0.05$. The jth box is erected over the interval $(2(j-1)/n, 2j/n]$, $j = 1, 2, \ldots, n-k = 38$, and encloses an area equal to the probability $P(L_{40}^{(2)} = j) = P(2L_{40}^{(2)}/40 = 2j/40)$. Only the first 23 boxes are visible as the probabilities become too small after $j = 23$.

Observe the anomalous behavior of the first box and the compensating discrepancy near 1. This surprising edge effect can be explained by a simple calculation: Since $\left\langle {s \atop 1} \right\rangle = 1$, for all integers s greater than or equal to 0, the

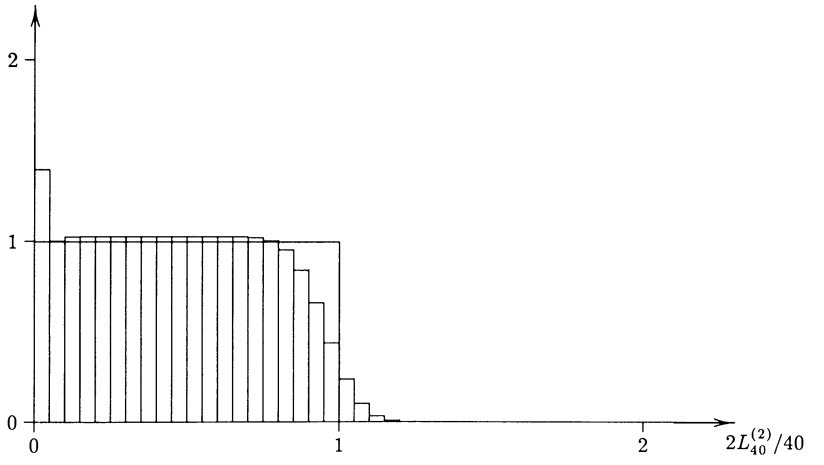

FIG. 2. *The exact distribution of* $2L_{40}^{(2)}/40$ *in comparison with* Beta(1, 1).

exact distribution of $2L_{40}^{(2)}/40$ at $j = 1$ yields

$$\begin{aligned} P\left(2L_{40}^{(2)}/40 = 2/40\right) &= \frac{1}{(n-1)(n-2)} \sum_{m=1}^{n-2} \frac{n-m}{(m-1)!} \\ &= \frac{1}{(n-1)(n-2)} \sum_{m=1}^{n-2} \frac{(n-1)-(m-1)}{(m-1)!} \\ &\approx \frac{1}{(n-1)(n-2)} \left[(n-1)e - e\right] \\ &\approx \frac{e}{n-1}. \end{aligned}$$

With $n = 40$, this probability is about 7% of the whole distribution. Thus the first box of the histogram has a 7% share of the distribution, and with width 0.05, the height of this box should be about 1.4, well above 1. This edge effect persists even for large n, since the height of the first box for large n will always be about $e/(n-1) \times n/2 \approx e/2 \approx 1.36$, well above 1. Of course, the width of the first box, $2/n$, tends to 0 as $n \to \infty$, so the limiting distribution is approached. A similar edge effect is observed for larger values of k.

REFERENCES

ATHREYA, K. B. and NEY, P. E. (1972). *Branching Processes*. Springer, New York.

FREEDMAN, D. (1965). Bernard Friedman's urn. *Ann. Math. Statist.* **36** 956–970.

FRIEDMAN, B. (1949). A simple urn model. *Comm. Pure Appl. Math.* **2** 59–70.

GASTWIRTH, J. (1977). A probability model of a pyramid scheme. *Amer. Statist.* **31** 79–82.

GASTWIRTH, J. and BHATTACHARYA, P. K. (1984). Two probability models of pyramids or chain letter schemes demonstrating that their promotional claims are unreliable. *Oper. Res.* **32** 527–536.

HALL, P. and HEYDE, C. C. (1980). *Martingale Limit Theory and Applications*. Academic, New York.

JOHNSON, N. and KOTZ, S. (1977). *Urn Models and Their Applications*. Wiley, New York.

KNUTH, D. E. (1973). *The Art of Computer Programming: Sorting and Searching* **3**. Addison-Wesley, Reading, Mass.

MOON, J. (1974). The distance between nodes in recursive trees. *London Math. Soc. Lecture Notes* **13** 125–132. Cambridge Univ. Press.

NA, H. and RAPOPORT, A. (1970). Distribution of nodes of a tree by degree. *Math. Biosci.* **6** 313–329.

NAJOCK, D. and HEYDE, C. C. (1982). On the number of terminal vertices in certain random trees with an application to stemma construction in philology. *J. Appl. Probab.* **19** 675–680.

SZYMAŃSKI, J. (1990). Branches in recursive trees. *J. Discrete Appl. Math.* To appear.

DEPARTMENT OF STATISTICS /
COMPUTER AND INFORMATION SYSTEMS
GEORGE WASHINGTON UNIVERSITY
WASHINGTON, D.C. 20052