LIMIT LAWS FOR LOCAL COUNTERS IN RANDOM BINARY SEARCH TREES

Luc Devroye
School of Computer Science
McGill University

Abstract. Limit laws for several quantities in random binary search trees that are related to the local shape of a tree around each node can be obtained very simply by applying central limit theorems for m-dependent random variables. Examples include: the number of leaves (L_n), the number of nodes with k descendants (k fixed), the number of nodes with no left child, the number of nodes with k left descendants. Some of these results can also be obtained via the theory of urn models, but the present method seems easier to apply.

Keywords and phrases. Binary search tree, data structures, probabilistic analysis, limit law, convergence, uniform random recursive trees, random trees.

CR Categories: 3.74, 5.25, 5.5.

Authors' address: School of Computer Science, McGill University, 3480 University Street, Montreal, Canada H3A 2K6. The authors' research was sponsored by NSERC Grant A3456 and FCAR Grant 90-ER-0291.

Introduction.

In this note, we consider a random binary search tree with n nodes obtained by inserting, in the standard manner, the values $\sigma_1, \dots, \sigma_n$ of a random permutation of $\{1, \dots, n\}$ into an initially empty tree. Equivalently, the search tree is obtained by inserting n i.i.d. uniform $[0,1]$ random variables $X_1, \dots, X_n$. Most shape-related quantities of the tree have been well–studied, including the expected depth and the exact distribution of the depth of X_n (Knuth, 1973; Lynch, 1965), the limit theory for the depth (Mahmoud and Pittel, 1984, Devroye, 1988), the first two moments of the internal path length (Sedgewick, 1983), the limit theory for the height of the tree (Pittel, 1984; Devroye, 1986, 1987), and various connections with the theory of random permutations (Sedgewick, 1983) and the theory of records (Devroye, 1988). Surveys of known results can be found in Vitter and Flajolet (1987) and Gonnet (1984).

The shape of the binary search tree is to some extent captured in quantities such as

L_n: the number of leaves;
O_n: the number of nodes with one child;
T_n: the number of nodes with two children;
R_n: the number of nodes with no left child;
V_{kn}: the number of nodes with k proper descendants.
L_{kn}: the number of nodes with k proper left descendants.

All of these describe the number of nodes having a certain "local" property. Several results are known about these quantities, e.g. in 1986, Mahmoud showed that $\mathbb{E}L_n \sim \mathbb{E}O_n \sim \mathbb{E}T_n \sim n/3$. The purpose of this note is to give a useful method of proving limit laws for all such "local" quantities. In the process, we will also gain insight into why Mahmoud's interesting result is true. Aldous (1990) gives a general methodolgy based upon urn models and branching processes for obtaining the first order behavior of the local quantities; his methods apply to a wide variety of trees; for the binary search tree, he has shown, among other things, that $V_{kn}/n \to 2/(k+2)(k+3)$ in probability as $n \to \infty$. We will give a short proof of this, and obtain the limit law for V_{kn} as well.

The general proof method.

It is convenient to think of the data in terms of pairs

$$(X_1, Y_1), \dots, (X_n, Y_n),$$

where the Y_i's are time stamps, which for the time being, can be defined by $Y_i \equiv i$. Thus, X_i is inserted before X_j if $Y_i < Y_j$. The data can also be reordered according to increasing X_i values: $X_{(1)} < \cdots < X_{(n)}$. In this case, we write

$$(X_{(1)}, Y_{(1)}), \dots, (X_{(n)}, Y_{(n)}).$$

We call a random variable N_n defined on a random binary search tree a **local counter** of order k if it can be written in the form

$$N_n = \sum_{i=1}^{n} f(Y_{(i-k)}, \ldots, Y_{(i)}, \ldots, Y_{(i+k)}),$$

where k is a fixed constant, $Y_{(i)} = 0$ if $i \le 0$ or $i > n$, and f is a $\{0,1\}$-valued function that is invariant under transformations of the Y_i's that keep the relative order of the Y's intact.

All the quantities introduced in the previous sections are local counters. For example, note that $X_{(i)}$ is a leaf if and only if $Y_{(i)}$ is larger than $Y_{(i-1)}$ and $Y_{(i+1)}$. Indeed, at the time of insertion of $X_{(i)}$, the tree consists of nodes with smaller time stamps than $Y_{(i)}$. The father of $X_{(i)}$ is the endpoint with the largest Y-value of the interval to which $X_{(i)}$ belongs in the partition of $[0,1]$ carved out by the first $i-1$ data points. Thus, we have the local counter representation

$$L_n = \sum_{i=1}^{n} I_{[Y_{(i)} > Y_{(i-1)}, Y_{(i)} > Y_{(i+1)}]}.$$

Similar representations exist for O_n, T_n and V_{kn}.

With local counters, the invariance allows us to replace the Y_i's by a sequence of i.i.d. uniform $[0,1]$ random variables; this in fact corresponds to introducing a (harmless) random permutation of the X_i's when we construct the binary search tree. Note, in particular, that $Y_{(1)}, \ldots, Y_{(n)}$ is itself an i.i.d. uniform $[0,1]$ sequence. Local counters have two key properties:

A. The i-th and j-th terms in the definition of N_n are independent whenever $|i-j| > 2k$.

B. The distribution of the i-th term is the same for all $i \in \{k+1, \ldots, n-k\}$. Thus, we have the representation $N_n = A_n + \sum_{i=k+1}^{n-k} Z_i$, where $0 \le A_n \le 2k$, and where the Z_i's are identically distributed and $2k$-dependent (a sequence of random variables Z_i is m-dependent if $(Z_1, \ldots, Z_i)$ is independent of the vector $(Z_j, \ldots)$ for any $j > i+m$). Observe that 0-dependence corresponds to independence.

Let $\mathcal{N}(0,\sigma^2)$ denote the normal distribution with mean 0 and variance σ^2. We will use a simple version of the central limit theorem for m-dependent stationary sequences due to Hoeffding and Robbins (1949):

LEMMA 1. *Let $Z_1, \ldots, Z_n, \ldots$ be a stationary sequence of random variables (i.e., for any k, the distribution of $(Z_i, \ldots, Z_{i+k})$ does not depend upon i), and let it also be m–dependent with m held fixed. Then, if $\mathbb{E}|Z_1|^3 < \infty$, the random variable*

$$\sum_{i=1}^{n} (Z_i - \mathbb{E}Z_i)/\sqrt{n} \longrightarrow \mathcal{N}(0,\sigma^2) \quad \textit{in distribution,}$$

where

$$\sigma^2 = \mathbb{V}(Z_1) + 2\sum_{i=2}^{m+1} (\mathbb{E}Z_1 Z_i - \mathbb{E}Z_1 \mathbb{E}Z_i).$$

The standard central limit theorem for independent (or 0-dependent) random variables is obtained as a special case. Subsequent generalizations of Lemma 1 were obtained by Brown (1971), Dvoretzky (1972), McLeish (1974), Ibragimov (1975), Chen (1978), Hall and Heyde (1980) and Bradley (1981), to name just a few. As a corollary, we see that if $\mathbb{E}Z_1 \neq 0$, then

$$\sum_{i=1}^{n} Z_i/n \to \mathbb{E}Z_1 \ \text{ in probability}$$

as $n \to \infty$. Lemma 1 and its corollary are directly applicable to local counters. We have

THEOREM 1. *Let N_n be a local counter for a random binary search tree, with fixed parameter k. Define $Z_i = f(U_i, \ldots, U_{i+2k})$, where $U_1, U_2, \ldots$ is a sequence of i.i.d. uniform $[0,1]$ random variables. Then*

$$(N_n - n\mathbb{E}Z_1)/\sqrt{n} \longrightarrow \mathcal{N}(0, \sigma^2) \quad \textit{in distribution,}$$

where

$$\sigma^2 = \mathbb{V}(Z_1) + 2 \sum_{i=2}^{2k+1} (\mathbb{E}Z_1 Z_i - \mathbb{E}Z_1 \mathbb{E}Z_i).$$

If $\mathbb{E}Z_1 \neq 0$, then $N_n/n \to \mathbb{E}Z_1$ in probability and in the mean as $n \to \infty$.

PROOF. We begin by recalling that $(Y_{(1)}, \ldots, Y_{(n)})$ is distributed as $(U_1, \ldots, U_n)$. Thus, in the notation of Theorem 1, the random variable $N_n - 2k$ is distributed as $\sum_{i=1}^{n-2k} Z_i$, and satisfies the conditions of Lemma 1. Thus,

$$(N_n - 2k - (n-2k)\mathbb{E}Z_1)/\sqrt{n} \longrightarrow \mathcal{N}(0, \sigma^2) \quad \text{in distribution.}$$

Here we used the fact that the Z_i's are $2k$-dependent. But

$$\left| \frac{N_n - n\mathbb{E}Z_1}{\sqrt{n}} - \frac{N_n - 2k - (n-2k)\mathbb{E}Z_1}{\sqrt{n}} \right| \leq \frac{4k}{\sqrt{n}} = o(1),$$

so that the first statement of Theorem 1 follows without work. The second statement follows from the first one. □

The number of leaves.

From Theorem 1, we obtain

THEOREM 2. *As $n \to \infty$, $(L_n - n/3)/\sqrt{n} \longrightarrow \mathcal{N}(0, 2/45)$ in distribution. An identical asymptotic result is valid for T_n. Also, $(O_n - n/3)/\sqrt{n} \longrightarrow \mathcal{N}(0, 8/45)$ in distribution.*

PROOF. Since L_n is a local counter with parameter $k = 1$, we have the representation (in distribution):

$$L_n = A_n + \sum_{i=2}^{n-1} Z_i,$$

where $0 \leq A_n \leq 2$, and

$$Z_i = I_{[Y_{(i)} > Y_{(i-1)}, Y_{(i)} > Y_{(i+1)}]}.$$

By Theorem 1, as $n \to \infty$, $(L_n - \mathbb{E}L_n)/\sqrt{n}$ has a limiting normal distribution with zero mean and variance

$$\sigma^2 = \mathbb{V}(Z_2) + 2\sum_{i=3}^{4}(\mathbb{E}Z_2 Z_i - \mathbb{E}Z_2\mathbb{E}Z_i).$$

We claim the following:

$$\mathbb{E}Z_2 = 1/3, \mathbb{V}(Z_2) = 2/9, \mathbb{E}Z_2Z_3 = 0, \mathbb{E}Z_2Z_4 = 2/15.$$

Thus,

$$\sigma^2 = 2/9 + 2(2/15 - 2/9) = 4/15 - 2/9 = 2/45.$$

The only possible difficulty is in the computation of $\mathbb{E}Z_2Z_4 = \mathbb{P}\{Y_{(2)} > Y_{(1)}, Y_{(2)} > Y_{(3)}, Y_{(4)} > Y_{(3)}, Y_{(4)} > Y_{(5)}\}$. We have five consecutive Y-values; these can be ordered in 5! ways. Of these, the desired configuration, in which the second and fourth values dominate their neighbors, occurs in 12+4 ways. The $12 = 2! \times 3!$ ways happen when the second and fourth values are one-two; the 4 ways occur when the Y-values are ordered 2-1-4-3-5, 2-1-5-3-4, 4-3-5-1-2 and 5-3-4-1-2. Thus, the probability is $16/120 = 2/15$.

The quantities L_n, O_n and T_n are closely related, since

$$L_n + O_n + T_n = n,$$

$$L_n = T_n + 1.$$

This implies that $O_n = n - 1 - 2L_n$, $T_n = L_n - 1$. Thus, $\mathbb{E}L_n \sim n/3$ implies the same thing for $\mathbb{E}O_n$ and $\mathbb{E}L_n$. Furthermore, $\mathbb{V}(O_n) \sim 4\mathbb{V}(L_n) \sim \mathbb{V}(T_n)$. Therefore, T_n follows the same limit laws as L_n, while $(O_n - n/3)/\sqrt{n}$ tends to a normal distribution with zero mean and variance $8/45$. □

REMARK. The moments of L_n can also be obtained with great ease. For example, exploiting symmetry, we have for $n \geq 2$,

$$\mathbb{E}L_n = 2\mathbb{P}\{Y_{(1)} > Y_{(2)}\} + (n-2)\mathbb{P}\{Y_{(2)} > Y_{(1)}, Y_{(2)} > Y_{(3)}\} = 1 + \frac{n-2}{3} = \frac{n+1}{3}. \;\square$$

Nodes with no left child.

Let R_n denote the number of nodes in a random binary search tree having no left child. Computations analogous to those of the previous sections show that we have the following:

THEOREM 3. *As $n \to \infty$, $(R_n - n/2)/\sqrt{n} \longrightarrow \mathcal{N}(0, 1/12)$ in distribution.*

PROOF. Using the (X_i, Y_i) representation of binary search trees introduced above, we see that $X_{(i)}$ has no left child if and only if either $i = 1$ or $i > 1$ and $Y_{(i)} > Y_{(i-1)}$. thus,

$$R_n = 1 + \sum_{i=2}^{n} I_{[Y_{(i)} > Y_{(i-1)}]}.$$

We have once again a representation in terms of a sum of $n - 1$ random variables that form a stationary 1-dependent sequence. Simple logic shows that $\mathbb{P}\{Y_{(2)} > Y_{(1)}\} = 1/2$ and that $\mathbb{P}\{Y_{(3)} > Y_{(2)} > Y_{(1)}\} = 1/6$. Thus, by Theorem 1, $(R_n - n/2)/\sqrt{n}$ is asymptotically normal with mean zero and variance $\sigma^2 = 1/4 + 2(1/6 - 1/4) = 1/12$.□

Leaves in uniform random recursive trees.

A uniform random recursive tree is an ordered tree in which node 1 is the root, and which is grown by adding node $n + 1$ simply by choosing its father uniformly and at random from among the n nodes $1, \ldots, n$ already present in the tree. Let M_n be the number of leaves in this tree. It is known that $M_n/n \to 1/2$ in probability and in the mean, and that $(M_n - n/2)/\sqrt{n} \longrightarrow \mathcal{N}(0, 1/12)$ in distribution (Najock and Heyde, 1982). We would simply like to point out that this result is also immediate from Theorem 2. Indeed, consider the oldestchild-nextsibling binary tree associated with the ordered tree (see Aho, Hopcroft and Ullman, 1983, for definitions). Choosing a random father for node $n+1$ is like picking a random external node in the binary tree with the proviso that the root's right external node is never picked. Thus, if we chop off the root of the associated binary tree, we obtain a random binary search tree on $n-1$ nodes. Now, M_n is equal to the number of nodes in the associated binary tree with no left child. This number is covered by Theorem 3. For other properties on uniform random recursive trees, see Dondajewski and Szymański (1982) and Na and Rapoport (1970).

The number of nodes with k descendants in a uniform random recursive tree is equal to the number of nodes in the associated random binary search tree with k left descendants. Let us denote by L_{kn} the number of nodes with k left descendants in a random binary search tree on n nodes. Then the following is true.

THEOREM 4. *Define* $p_k = 1/(k+2)(k+1)$. *Then*

$$\frac{L_{kn}}{n} \longrightarrow p_k \quad \textit{in probability}$$

and

$$\frac{L_{kn} - np_k}{\sqrt{n}} \longrightarrow \mathcal{N}(0, \sigma_k^2) \quad \textit{in distribution}$$

as $n \to \infty$, *where*

$$\sigma_k^2 = p_k(1-p_k) - 2(k+1)p_k^2 + 2\rho_k$$

and

$$\rho_k = \frac{1}{(2k+3)(2k+2)(k+1)} \ .$$

PROOF. We have the representation

$$L_{kn} = \sum_{i=1}^{n} Z_i \ ,$$

where

$$Z_i = \begin{cases} 0 \ , & i \le k; \\ I_{[Y_{(i)} < \min(Y_{(i-1)}, \ldots, Y_{(i-k)})]} \ , & i = k+1; \\ I_{[Y_{(i-k-1)} < Y_{(i)} < \min(Y_{(i-1)}, \ldots, Y_{(i-k)})]} \ , & i > k+1. \end{cases}$$

A trivial argument shows that $\mathbb{E}Z_i = k!/(k+2)! = p_k$ for $i > k+1$. Also, for $n \ge j > i > k+1$, we have

$$\mathbb{E}Z_i Z_j = \begin{cases} 0 \ , & j > i > j-k-1; \\ \rho_k \ , & i = j-k-1; \\ p_k^2 \ , & i < j-k-1. \end{cases}$$

To see this, note that

$$\begin{aligned} \rho_k = \mathbb{P}\{ & Y_{(i-k-1)} < Y_{(i)} < \min(Y_{(i-1)}, \ldots, Y_{(i-k)}) \ , \\ & Y_{(i)} < Y_{(i+k+1)} < \min(Y_{(i+k)}, \ldots, Y_{(i+1)})\} \ . \end{aligned}$$

There are $(2k+3)!$ ways of permuting $Y_{(i-k-1)}, \ldots, Y_{(i+k+1)}$. To compute ρ_k is to count the number of permutations yielding $Y_{(i-k-1)} < Y_{(i)} < Y_{(i+k+1)}$, while at the same time $Y_{(i)} < \min(Y_{(i-k)}, \ldots, Y_{(i-1)})$ and $Y_{(i+k+1)} < \min(Y_{(i+1)}, \ldots, Y_{(i+k)})$. The arguments of the two minima can be permuted in $k!$ ways each. The total number of desired permutations is

$$k!^2 \binom{2k+1}{k} \ .$$

Thus,

$$\rho_k = \frac{1}{(2k+3)(2k+2)(k+1)} \ .$$

By Lemma 1,

$$L_{kn}/n \longrightarrow p_k \quad \text{in probability}$$

and

$$\frac{L_{kn} - np_k}{\sqrt{n}} \longrightarrow \mathcal{N}(0, \sigma_k^2) \quad \text{in distribution,}$$

where

$$\begin{aligned}\sigma_k^2 &= \mathbb{V} Z_{k+2} + 2 \sum_{j=k+3}^{2k+3} (\mathbb{E} Z_{k+2} Z_j - \mathbb{E} Z_{k+2} \mathbb{E} Z_j) \\ &= p_k(1-p_k) + 2\mathbb{E} Z_{k+2} Z_{2k+3} - 2(k+1)p_k^2 \\ &= p_k(1-p_k) + 2\rho_k - 2(k+1)p_k^2 \ .\end{aligned}$$

Here we took $i = k+2$ only to rid ourselves of the boundary effects. This concludes the proof of the Theorem. □

REMARK. In the special case $k = 0$, we have $p_0 = 1/2$, $\rho_k = 1/6$, $\sigma_k^2 = 2\rho_k - p_k^2 = 1/12$. In the case $k = 1$, we have $p_k = 1/6$, $\rho_k = 1/40$, and $\sigma_k^2 = 7/90$. □

Nodes with exactly k descendants.

Let k be fixed, independent of n. Simple considerations show that V_{kn}, the number of nodes with precisely k descendants, is indeed a local counter. Note that all the proper descendants of a node $X_{(i)}$ are found by finding the largest $0 \le j < i$ with $Y_{(j)} < Y_{(i)}$, and the smallest l greater than i and no more than n such that $Y_{(l)} < Y_{(i)}$. All the nodes $X_{(j+1)}, \ldots, X_{(l-1)}$, $X_{(i)}$ excluded, are proper descendants of $X_{(i)}$. Thus, to decide whether $X_{(i)}$ has exactly k descendants, it suffices to look at $Y_{(i-k-1)}, \ldots, Y_{(i+k+1)}$, so that V_{kn} is a local counter with parameter $k+1$. Theorem 2 above implies the following:

THEOREM 5. *Let V_{kn} be the number of nodes with k descendants. Then*

$$\frac{V_{kn}}{n} \longrightarrow p_k \stackrel{\text{def}}{=} \frac{2}{(k+3)(k+2)} \quad \textit{in probability}$$

and

$$\frac{V_{kn} - np_k}{\sqrt{n}} \longrightarrow \mathcal{N}(0, \sigma_k^2) \quad \textit{in distribution}$$

as $n \to \infty$, where

$$\sigma_k^2 \stackrel{\text{def}}{=} p_k(1-p_k) + 2(k+1)^2 \rho_k - 2(k+2)p_k^2$$

and

$$\rho_k \stackrel{\text{def}}{=} \frac{5k+8}{(k+1)^2(k+2)^2(2k+5)(2k+3)} \ .$$

REMARK. The first part of this theorem is also implicit in Aldous (1990). □

REMARK. When $k = 0$, we get $p_0 = 1/3$, $\rho_0 = 2/15$, $\sigma_0^2 = 2/45$. For $k = 1$, we obtain $p_1 = 1/6$, $\rho_1 = 13/1260$ and $\sigma_1^2 = 23/420$. □

PROOF. We have the representation

$$V_{kn} = \sum_{i=1}^{n} Z_i \ ,$$

where

$$Z_i \stackrel{\text{def}}{=} \sum_{j=0}^{k} Z_i(j, k-j) \ ,$$

and $Z_i(j,l)$ is the indicator of the event that $(X_{(i)}, Y_{(i)})$ has j left descendants and l right descendants. Assume throughout that $1 \le i-k-1$, $i+k+1 \le n$ when we discuss Z_i. The values $Z_1, \ldots, Z_{k+1}$ and $Z_{n-k}, \ldots, Z_n$ are all zero or one, and affect V_{kn} jointly by at most $2k+2$ (which is a constant). We also have the representation, for $1 \le i-k-1$, $i+l+1 \le n$:

$$\begin{aligned} Z_i(j,l) = {} & I_{[Y_{(i-j-1)} < Y_{(i)} < \min(Y_{(i-1)}, \ldots, Y_{(i-j)})]} \\ & \times I_{[Y_{(i+l+1)} < Y_{(i)} < \min(Y_{(i+1)}, \ldots, Y_{(i+l)})]} \ . \end{aligned}$$

A simple argument shows that for i, j, l as restricted above,

$$\mathbb{E} Z_i(j,l) = \frac{2(j+l)!}{(j+l+3)!} = \frac{2}{(j+l+3)(j+l+2)(j+l+1)} \ .$$

Thus, for $1 \le i-k-1$, $i+k+1 \le n$,

$$\mathbb{E} Z_i(j, k-j) = \frac{2}{(k+3)(k+2)(k+1)} \stackrel{\text{def}}{=} q_k \ ,$$

and

$$\mathbb{E} Z_i = \sum_{j=0}^{k} \mathbb{E} Z_i(j, k-j) = \sum_{j=0}^{k} q_k = \frac{2}{(k+3)(k+2)} \ .$$

It is clear that V_{kn} is a local counter for a random binary search tree, so we may apply Theorem 1. To do so, we need to study $\mathbb{E} Z_i Z_{i+r}$, where $1 \le i-k-1$, $i+r+k+1 \le n$, $1 \le i \le n$, $1 \le r$. For $0 \le j \le k$, $0 \le l \le n$, we claim the following:

$$\mathbb{E} Z_i(j,k-j) Z_{i+r}(l,k-l) = \begin{cases} 0 \ , & \text{if } r < k-j+l+2; \\ \rho_k \ , & \text{if } r = k-j+l+2; \\ \mathbb{E} Z_i(j,k-j) \mathbb{E} Z_{i+r}(l,k-l) \ , & \text{if } r > k-j+l+2, \end{cases}$$

where

$$\begin{aligned} \rho_k & \stackrel{\text{def}}{=} \frac{k!^2}{(2k+5)!} \left\{ \binom{2k+4}{k+2} + 2\binom{2k+3}{k+2} + 2\binom{2k+2}{k+1} \right\} \\ & = \frac{5k+8}{(k+1)^2(k+2)^2(2k+5)(2k+3)} \ . \end{aligned}$$

The last expression is obtained by noting that of the $(2k+5)!$ possible permutations of $Y_{(i-j-1)}, \ldots, Y_{(i+r+k-l+1)}$, with $r = k-j+l+2$, only $\rho_k (2k+5)!$ are such that $Z_i(j,k-j) Z_{i+r}(l,k-l) = 1$. The three terms in the expression of ρ_k are obtained by considering

A. $Y_{(i+k-j+1)}$ is smaller than both $Y_{(i-j-1)}$ and $Y_{(i+r+k-l+1)}$.

B. $Y_{(i+k-j+1)}$ is smaller than one of $Y_{(i-j-1)}$ and $Y_{(i+r+k-l+1)}$.

C. $Y_{(i+k-j+1)}$ is larger than both $Y_{(i-j-1)}$ and $Y_{(i+r+k-l+1)}$.

If $r > 2k+2$, then Z_i and Z_{i+r} are independent. Thus, we need only consider the case $1 \leq r \leq 2k+2$. Let L, J be independent random variables uniformly distributed on $\{0, \ldots, k\}$.

$$\begin{aligned}
\mathbb{E}Z_i Z_{i+r} &= \mathbb{E}\left\{\sum_{j=0}^{k} Z_i(j, k-j) \sum_{l=0}^{k} Z_{i+r}(l, k-l)\right\} \\
&= \sum_{j=0}^{k}\sum_{l=0}^{k} \rho_k I_{[r = k-j+l+2]} + \sum_{j=0}^{k}\sum_{l=0}^{k} q_k^2 I_{[r > k-j+l+2]} \\
&= (k+1)^2 \rho_k \mathbb{P}\{r = k-J+L+2\} + (k+1)^2 q_k^2 \mathbb{P}\{r > k-J+L+2\} \ .
\end{aligned}$$

Summing this gives

$$\begin{aligned}
&\sum_{r=1}^{2k+2} (\mathbb{E}Z_i Z_{i+r} - \mathbb{E}Z_i \mathbb{E}Z_{i+r}) \\
&\quad = (k+1)^2 \rho_k \sum_{r=1}^{2k+2} \mathbb{P}\{J-L = k+2-r\} \\
&\qquad + (k+1)^2 q_k^2 \sum_{r=1}^{2k+2} \mathbb{P}\{J-L > k+2-r\} - (2k+2)p_k^2 \\
&\quad = (k+1)^2 \rho_k + p_k^2 \sum_{r=1}^{k+1} \mathbb{P}\{J-L > k+2-r\} \\
&\qquad + p_k^2 \sum_{r=k+2}^{2k+2} \mathbb{P}\{J-L > k+2-r\} - (2k+2)p_k^2 \\
&\quad = (k+1)^2 \rho_k + p_k^2 \sum_{r=1}^{k+1} \mathbb{P}\{J-L > r\} \\
&\qquad + p_k^2 \sum_{r=0}^{k} \mathbb{P}\{J-L > -r\} - (2k+2)p_k^2 \\
&\quad = (k+1)^2 \rho_k + p_k^2 \sum_{r=2}^{k+2} \mathbb{P}\{J-L \geq r\} \\
&\qquad + p_k^2 \sum_{r=0}^{k} (1 - \mathbb{P}\{J-L \geq r\}) - (2k+2)p_k^2 \\
&\quad = (k+1)^2 \rho_k - p_k^2 \sum_{r=0}^{1} \mathbb{P}\{J-L \geq r\} + p_k^2(k+1) - (2k+2)p_k^2 \\
&\quad = (k+1)^2 \rho_k - p_k^2 + p_k^2(k+1) - (2k+2)p_k^2 \\
&\quad = (k+1)^2 \rho_k - (k+2)p_k^2 \ .
\end{aligned}$$

By Lemma 1, $V_{kn}/n \to p_k$ in probability as $n \to \infty$ and

$$\frac{V_{kn} - np_k}{\sqrt{n}} \longrightarrow \mathcal{N}(0, \sigma_k^2) \quad \text{in distribution,}$$

where

$$\sigma_k^2 = p_k(1-p_k) + 2(k+1)^2\rho_k - 2(k+2)p_k^2 \ . \ \square$$

Urn models.

The limit law for L_n can be obtained by several methods. The method presented above is simple and didactical. Another method uses the properties of Pólya-Eggenberger urn models, which have been suggested for the analysis of search trees by Poblete and Munro (1985). Bagchi and Pal (1985) developed a limit law for general urn models and applied it in the analysis of random 2-3 trees.

In a binary search tree with n nodes, let W_n be the number of external nodes with another sibling external node, and let B_n count the remaining external nodes. Clearly, $W_n + B_n = n+1$, $W_0 = 0$ and $B_0 = 1$. When a random binary search tree is grown, each external node is picked with equal probability (see e.g. Knuth and Schonhage, 1978). Thus, upon insertion of node $n+1$, we have:

$$(W_{n+1}, B_{n+1}) = (W_n, B_n) + \begin{cases} (0,1) & \text{with probability } \frac{W_n}{W_n+B_n}; \\ (2,-1) & \text{with probability } \frac{B_n}{W_n+B_n}. \end{cases}$$

This is known as a generalized Pólya-Eggenberger urn model. The model is defined by the matrix

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 2 & -1 \end{pmatrix}$$

For general values of a, b, c, d, the asymptotic behavior of W_n is governed by the following Lemma (Bagchi and Pal, 1985) (for a special case, see e.g. Athreya and Ney, 1972):

Lemma 2. *Consider an urn model in which* $a + b = c + d \stackrel{\text{def}}{=} s \geq 1$, $W_0 + B_0 \geq 1$, $0 \leq W_0$, $0 \leq B_0$, $a \neq c$, $b, c > 0$, $a - c \leq s/2$, *and, if* $a < 0$, *then* a *divides both* c *and* W_0, *and if* $d < 0$, *then* d *divides both* b *and* B_0. *Then*

$$\frac{W_n}{W_n + B_n} \to \frac{c}{b+c} \quad \text{almost surely,}$$

and

$$\frac{W_n - \mathbb{E}W_n}{\sqrt{n}} \longrightarrow \mathcal{N}(0, \sigma^2) \quad \textit{in distribution,}$$

where

$$\sigma^2 = \frac{bc}{(b+c)^2} \frac{(s-b-c)^2}{2b+2c-s}.$$

In our case, $\sigma^2 = 8/45$. Since $L_n = W_n/2$, the variance of L_n is one fourth that of W_n, so Theorem 1 follows immediately from Lemma 2 as well. Additionally, Lemma 2 implies that $W_n/n \to 1/3$ almost surely.

References.

A. V. Aho, J. E. Hopcroft, and J. D. Ullman, *Data Structures and Algorithms*, Addison-Wesley, Reading, Mass., 1983.

D. Aldous, "Asymptotic fringe distributions for general families of random trees," Technical Report, Department of Statistics, University of California at Berkeley, 1990.

K. B. Athreya and P. E. Ney, *Branching Processes*, Springer Verlag, Berlin, 1972.

A. Bagchi and A. K. Pal, "Asymptotic normality in the generalized Polya-Eggenberger unrn model, with an application to computer data structures," *SIAM Journal on Algeraic and Discrete Methods*, vol. 6, pp. 394–405, 1985.

R. C. Bradley, "Central limit theorems under weak dependence," *Journal of Multivariate Analysis*, vol. 11, pp. 1–16, 1981.

B. M. Brown, "Martingale central limit theorems," *Annals of Mathematical Statistics*, vol. 42, pp. 59–66, 1971.

L. H. Y. Chen, "Two central limit problems for dependent random variables," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 43, pp. 223–243, 1978.

L. Devroye, "A note on the height of binary search trees," *Journal of the ACM*, vol. 33, pp. 489–498, 1986.

L. Devroye, "Branching processes in the analysis of the heights of trees," *Acta Informatica*, vol. 24, pp. 277–298, 1987.

L. Devroye, "Applications of the theory of records in the study of random trees," *Acta Informatica*, vol. 26, pp. 123–130, 1988.

M. Dondajewski and J. Szymański, "On the distribution of vertex-degrees in a strata of a random recursive tree," *Bulletin de l'Académie Polonaise des Sciences, Série des Sciences Mathématiques*, vol. 30(5-6), pp. 205–209, 1982.

A. Dvoretzky, "Central limit theorems for dependent random variables," in: *Proceedings of the Sizth Berkeley symposium on Mathematical statistics and Probability*, pp. 513–535, University of California Press, 1972.

G. H. Gonnet, *A Handbook of Algorithms and Data Structures*, Addison-Wesley, Reading, Mass., 1984.

P. Hall and C. C. Heyde, *Martingale Limit Theory and its Applications*, Academic Press, New York, 1980.

W. Hoeffding and H. Robbins, "The central limit theorem for dependent random variables," *Annals of Mathematical Statistics*, vol. 20, pp. 773–780, 1949.

I. A. Ibragimov, "A note on the central limit theorem for dependent random variables," *Theory of Probability and its Applications*, vol. 20, pp. 135–141, 1975.

D. E. Knuth, *The Art of Computer Programming, Vol. 1: Fundamental Algorithms*, Addison-Wesley, Reading, Mass., 1973. 2nd Ed.

D. E. Knuth and A. Schonhage, "The expected linearity of a simple equivalence algorithm," *Theoretical Computer Science*, vol. 6, pp. 281–315, 1978.

W. C. Lynch, "More combinatorial problems on certain trees," *Computer Journal*, vol. 7, pp. 299–302, 1965.

H. M. Mahmoud, "The expected distribution of degrees in random binary search trees," *The Computer Journal*, vol. 29, pp. 36–37, 1986.

H. Mahmoud and B. Pittel, "On the most probable shape of a search tree grown from a random permutation," *SIAM Journal on Algebraic and Discrete Methods*, vol. 5, pp. 69–81, 1984.

D. L. McLeish, "Dependent central limit theorems and invariance principles," *Annals of Probability*, vol. 2, pp. 620–628, 1974.

H. S. Na and A. Rapoport, "Distribution of nodes of a tree by degree," *Mathematical Biosciences*, vol. 6, pp. 313–329, 1970.

D. Najock and C. C. Heyde, "On the number of terminal vertices in certain random trees with an application to stemma construction in philology," *Journal of Applied Probability*, vol. 19, pp. 675–680, 1982.

B. Pittel, "On growing random binary trees," *Journal of Mathematical Analysis and Applications*, vol. 103, pp. 461–480, 1984.

P. V. Poblete and J. I. Munro, "The analysis of a fringe heuristic for binary search trees," *Journal of Algorithms*, vol. 6, pp. 336–350, 1985.

R. Sedgewick, "Mathematical analysis of combinatorial algorithms," in: *Probability Theory and Computer Science*, pp. 123–205, ed. G. Louchard and G. Latouche, Academic Press, London, 1983.

J. S. Vitter and P. Flajolet, “Average-case analysis of algorithms and data structures,” Technical Report CS-87-20, Department of Computer Science, Brown University, Providence, RI, 1987.