Exploring Werner Krandick's Binary Tree Jump Statistics

Shalosh B. EKHAD and Doron ZEILBERGER

Preface

Fabrice Rouillier and Paul Zimmermann [RZ] proposed a unified and very efficient algorithm for finding real roots of univariate polynomials based on the good-old *Descartes' rule of signs*. It improved previous algorithms due to George Collins and Alkiviadis G. Akritas, Jeremy Johnson, and Werner Krandick (see [RZ] for references). One reason it was so efficient was that in the process, the algorithm constructs a certain binary tree, that is traversed in depth-first-search and whenever there is a "jump" it is expensive. It turned out, that on average, there are not so many jumps, explaining the efficiently.

This motivated Werner Krandick [K] to find explicit expressions for the expectations of the statistics "number of jumps" and "sum of the jump-distances" (see below for the exact definitions). Using clever but ad hoc human reasoning, he found that they were $\frac{n-1}{2}$ and $\frac{n(n-1)}{n+2}$ respectively (Theorem 11 of [K]).

Here we show the power of symbolic computation and experimental mathematics to do much more. In particular, explicit expressions for the *variances* (namely $\frac{n^2-1}{8n-4}$ and $\frac{2n\left(2n^2-n-1\right)}{n^3+7n^2+16n+12}$ respectively). Better still, we will derive explicit expressions for the weight-enumerators of the set of full binary trees according to these statistics (and the number of internal vertices) from which the expectation, variance, and any number of higher moments can be easily deduced. Everything is implemented in the Maple package `Krandick.txt`, available from the front of this article:

`https://sites.math.rutgers.edu/~zeilberg/mamarim/mamarimhtml/krandick.html` .

But first *definitions*.

Definitions

• **A (full) binary tree** is an unlabeled ordered tree, where every vertex has either 0 children (and then it is called a *leaf*) or 2 children (and then it is called an *internal vertex*). A good way to define these creatures is *recursively*. A binary tree has either 0 internal vertices (i.e. it only consists of the root), or else the root has a *left son* that is the root of a binary tree T_L and a *right son*, that is the root of a binary tree T_R.

In symbols: either $T = .$ or $T = [T_L, T_R]$.

• Let $V(T)$ be the number of internal vertices of the binary tree T. Note that it may be defined recursively by:

$$V(.) = 0 \quad , \quad V([T_L, T_R]) = V(T_L) + V(T_R) + 1 \quad ,$$

since by removing the root, you lose an internal vertex.

• J(T) denotes the *number* of “jumps” when you traverse it in *depth-first-search* (see [K]). It is best defined *recursively* as follows

$$J(.) = 0$$

and

$$J([T_L, T_R]) = \begin{cases} J(T_R), & \text{if } T_L = . \quad ; \\ J(T_L) + J(T_R) + 1, & \text{otherwise.} \end{cases}$$

We also need an auxiliary statistic, the *depth of the rightmost leaf.* It may be defined recursively as follows.

$$D(.) = 0 \quad ,$$

and

$$D([T_L, T_R]) = 1 + D(T_R) \quad .$$

Another statistic studied in [K] is the *sum of jump distances.* It may be defined recursively as follows:

$$JD(.) = 0 \quad ,$$

and

$$JD([T_L, T_R]) = JD(T_L) + JD(T_R) + D(T_L) \quad .$$

As noticed in [K], it is readily seen that $JD(T) + D(T) = V(T)$, so in some sense, as we will see soon, the statistic JD is ‘easier’ than J.

Theorems

Theorem **0**: Let $\mathcal{B}$ be the (infinite) set of *all* binary trees, and let $f(x)$ be its *weight-enumerator* according to the weight W_0 defined by: $W_0(T) := x^{V(T)}$. Then $f(x) := W_0(\mathcal{B})$, a certain *formal power series* in x with *integer coefficients*, satisfies the quadratic equation

$$f(x) = 1 \, + \, x\, f(x)^2 \quad .$$

Proof: A binary tree is either trivial, with zero internal vertices, whose weight is $x^0 = 1$ explaining the ‘1’ on the right side of the equation, or it has a left tree and right tree, the x in front of the second term on the right is because when you remove the root you lose an internal vertex, and T_L and T_R range all over $\mathcal{B}$, and since $x^{a+b} = x^a x^b$, we have it.

Comment: Solving the quadratic equation gives

$$f(x) = \frac{1 - \sqrt{1-4x}}{2x} \quad ,$$

that thanks to Isaac Newton's binomial theorem, implies that the number of binary trees with n internal vertices, let's call it b_n, is given by the good old Catalan numbers

$$b_n = \frac{(2n)!}{n!(n+1)!} \quad .$$

See $[CDZ]$ for many other proofs of this result. Also see [S] for many other combinatorial objects counted by the Catalan numbers.

Theorem 1: Let $\mathcal{B}$ be the (infinite) set of *all* binary trees. Define a (tri-variate) weight, $W_1(T) := x^{V(T)}t^{D(T)}q^{J(T)}$, and let $F(x,t,q) := W_1(\mathcal{B})$, a certain formal power series in x with coefficients that are polynomials of t and q. $F(x,t,q)$ satisfies the following functional equation:

$$F(x,t,q) = 1 \,+\, x\,t\,F(x,0,q)\,F(x,t,q) \,+\, xtq\,(F(x,1,q) - F(x,0,q))\,\,F(x,t,q) \quad .$$

Proof: Follows easily from the recursive definitions of $V(T)$, $J(T)$, and $D(T)$.

Theorem 2: An explicit expression, in terms of 'radicals', for $F(x,t,q)$ is

$$F(x,t,q) \;=\; -\frac{-qtx + tx + \sqrt{q^2t^2x^2 - 2q\,t^2x^2 - 2q\,t^2x + t^2x^2 - 2t^2x + t^2} + t - 2}{2\,(qtx + t^2x - tx - t + 1)} \quad .$$

Proof: Let the right side be $G(x,t,q)$. We claim that

$$G(x,t,q) - (1 \,+\, x\,t\,G(x,0,q)\,G(x,t,q) \,+\, xtq\cdot(G(x,1,q) - G(x,0,q))\,\,G(x,t,q)) \;=\; 0 \quad ,$$

(check!). The theorem follows from the obvious **uniqueness** of the solution of this functional equation (in the ring of formal power series in x,t,q).

Secret from Kitchen: There is a sophisticated method (that we dislike!) called the *kernel method*, and presumably it could be done that way. But a much better way, is to *hope* that in addition to the functional equation, mixing $F(x,t,q)$ and $F(x,0,q)$ and $F(x,1,q)$, it also satisfies a *pure* quadratic equation with coefficients that are polynomials in x,t. So, using the combinatorially derived functional equation, we cranked out sufficiently many terms and *guessed* such an equation. We found it! Of course, so far this is *only* a guess. Then we asked Maple to `solve` it in radicals. This is still a guess. But **once conjectured** it is a *routine verification*, that Maple kindly did for us. See procedure `ProveJxtq()` in our Maple package `Krandick.txt`.

Indeed, if you downloaded `Krandick.txt` to your own laptop (that has Maple), please type:

`ProveJxtq();` ,

and before you know it you would get

`true`.

Note that we needed the variable t, corresponding to the 'depth of the rightmost leaf', in order to be able to set-up the functional equation, but we are really not interested in it! It is only a *stepping stone*, a *catalytic variable*. At the *end of the day*, we can plug-in $t = 1$ and get the following theorem.

Theorem 3: Let $\mathcal{B}$ be the (infinite) set of *all* binary trees. Define a (bivariate) weight, $W_2(T) := x^{V(T)}q^{J(T)}$, and let $H(x,q) := W_2(\mathcal{B})$, a certain formal power series in x with coefficients that are polynomials of q. We have:

$$H(x,q) = -\frac{-qx + \sqrt{q^2x^2 - 2q\,x^2 - 2qx + x^2 - 2x + 1} + x - 1}{2qx} \quad .$$

Theorem 4: Let $\mathcal{B}$ be the (infinite) set of *all* binary trees. Define a (bivariate) weight, $W_3(T) := x^{V(T)}t^{D(T)}$, and let $J(x,t) := W_3(\mathcal{B})$, a certain formal power series in x with coefficients that are polynomials of t. We have the following functional equation:

$$J(x,t) = 1 \,+\, x\,t\,J(x,1)J(x,t) \quad .$$

Proof: A member of $\mathcal{B}$ is either the singleton tree, '.', or else can be written as $T = [T_L, T_R]$. Since $V(T) = V(T_L) + V(T_R) + 1$, and $D(T) = D(T_R) + 1$, the equation follows (the left tree T_L does not contribute to the t part, so the variable t is set to 1).

Solving for $J(x,t)$ gives that it equals $1/(1 - xtJ(x,1))$. But $J(x,1)$ is nothing but our good old $f(x)$, the generating function for the Catalan numbers.

So we have

Theorem 5: An explicit expression for $J(x,t)$ is

$$J(x,t) \;=\; \frac{2}{t\sqrt{1-4x} - t + 2} \quad .$$

Theorem 6: Let $\mathcal{B}$ be the (infinite) set of *all* binary trees. Define a (bivariate) weight, $W_4(T) := x^{V(T)}q^{JD(T)}$, and let $K(x,q) := W_4(\mathcal{B})$, a certain formal power series in x with coefficients that are polynomials of q. We have the following explicit expression

$$K(x,q) \;=\; \frac{2q}{\sqrt{-4qx+1} - 1 + 2q} \quad .$$

Proof: We noticed above that $D(T) + JD(T) = V(T)$, hence $K(x,q) = J(qx, \frac{1}{q})$, and Theorem 6 follows from Theorem 5.

Moments of The Jump Statistics

The weight-enumerator contains **all** the information needed for *all* the moments.

In particular The generating function of the quantity

Sum of the 'number of jumps'

over all binary trees with n internal vertices, what Krandick [K] denoted by j_n, is the coefficient of x^n in $\frac{\partial}{\partial q} H(x,q)|_{q=1}$, that implies that the expected number of jumps is j_n/b_n, that happens to be $(n-1)/2$.

More generally, the generating function for the quantity

sum of the 'r^{th}*-power of the number of jumps'*

over all binary trees with n internal vertices, is

the coefficient of x^n in $(q\frac{\partial}{\partial q})^r H(x,q)|_{q=1}$.

Calling this quantity $j_n^{(r)}$, the r-th moment is $j_n^{(r)}/b_n$.

From the usual moments, one easily derives the *moments about the mean*, in particular the *variance*. Once we have explicit expressions for the moments about the mean (for as many as we desire), we get the *scaled moments* and then we can take the limit as n goes to ∞. To our pleasant surprise these (at least up to the 10-th moment) coincide with those of the normal distribution, 0 for odd moments, and $\frac{(2r)!}{2^r r!}$ for the $2r^{th}$ moment. This indicates that Krandick's jump statistics is most probably *asymptotically normal.* Can you prove it?

It is (probably) not hard to prove that the moments, and hence the moments about the mean, are **rational functions** of n, and one can easily bound the degrees of the numerator and denominators. So why not crank out many 'data values' and fit them into rational functions? That's exactly what we did. Notice that it is irrelevant whether we have an *a priori* proof that these are rational functions. Once *conjectured* it is a routine (rigorous!) verification.

So we have the following experimentally derived, but fully *rigorizable* theorems.

Theorem 7.1: (first proved in [K]) The expected 'number of jumps' among all binary trees with n internal vertices is

$$\frac{n-1}{2} \quad .$$

Theorem 7.2: The variance of the 'number of jumps' statistic among all binary trees with n internal vertices is

$$\frac{n^2-1}{8n-4} \quad .$$

Theorem 7.3: The kurtosis (aka 'scaled fourth moment-about-the-mean') of the 'number of jumps' statistic among all binary trees with n internal vertices is

$$\frac{6n^3 - 11n^2 - 2n + 3}{2n^3 - 3n^2 - 2n + 3} \quad .$$

(Note that it tends to 3, as it should).

Theorem 7.4: The 6^{th} scaled moment-about-the-mean of the 'number of jumps' statistic among all binary trees with n internal vertices is

$$\frac{60n^6 - 300n^5 + 391n^4 - 20n^3 - 82n^2 - 16n + 15}{4n^6 - 16n^5 + 7n^4 + 32n^3 - 26n^2 - 16n + 15} \quad ,$$

note that it tends to $1 \cdot 3 \cdot 5 = 15$, as it should).

Theorem 7.5: The scaled 8^{th} moment-about-the-mean of the 'number of jumps' statistic among all binary trees with n internal vertices is

$$\frac{840n^9 - 7980n^8 + 27006n^7 - 38933n^6 + 23070n^5 - 6937n^4 + 3178n^3 - 1167n^2 - 142n + 105}{8n^9 - 60n^8 + 118n^7 + 75n^6 - 402n^5 + 135n^4 + 418n^3 - 255n^2 - 142n + 105} \quad .$$

(Note that it tends to $1 \cdot 3 \cdot 5 \cdot 7 = 105$, as it should).

For the explicit expression for the tenth scaled moment about the mean, look at the output file:

`https://sites.math.rutgers.edu/~zeilberg/tokhniot/oKrandick1.txt` .

Moments of The Sum of Jump Distances Statistic

This one is even more **concentrated about the mean**, since as will see below, the variance tends to a constant.

Using our guessing methodology we have the following theorems.

Theorem 8.1: (first proved in [K]) The expected 'sum of jump distances' among all binary trees with n internal vertices is

$$\frac{n\,(n-1)}{n+2} \quad .$$

Theorem 8.2: The variance of the statistic 'sum of jump distances' among all binary trees with n internal vertices is

$$\frac{2n\left(2n^2 - n - 1\right)}{n^3 + 7n^2 + 16n + 12} \quad ,$$

(note that it converges to 4, hence the standard-deviation converges to 2).

Theorem 8.3: The *skewness* (aka *scaled third moment-about-the-mean*) of the statistic 'sum of jump distances' among all binary trees with n internal vertices is

$$\frac{3\sqrt{2}\sqrt{\frac{(n^3-n^2-8n+12)n}{2n^4+15n^3+23n^2-24n-16}}}{2} \quad ,$$

(note that it converges to $\frac{3}{2}$). In particular, this statistic is **not** asymptotically normal, since for the latter to be true it should have been 0).

Theorem 8.4: The kurtosis of the statistic 'sum of jump distances' among all binary trees with n internal vertices is

$$\frac{25n^5+58n^4-45n^3-34n^2-172n-48}{2\,(2n^4+17n^3+30n^2-29n-20)\,n} \quad ,$$

(note that it converges to $\frac{25}{4}$).

For explicit expressions for the fifth through the tenth scaled moments about the mean, look at the output file

`https://sites.math.rutgers.edu/~zeilberg/tokhniot/oKrandick2.txt` .

Conclusion

Werner Krandick used pure human cleverness to find explicit expressions for the *first* moments of the jump statistics that he studied. But using **symbolic computation** and **experimental mathematics**, one can go much further. *We believe that this is the way to go.*

References

[CDZ] Shaoshi Chen, Robert Dougherty-Bliss, and Doron Zeilberger, *C_4 proofs that the number of binary trees with n+1 leaves is given by the Catalan number C_n*, in preparation.

[K] Werner Krandick, *Trees and jumps and real roots*, Journal of Computational and Applied Mathematics **162** (2004), 51-55.
`https://sites.math.rutgers.edu/~zeilberg/akherim/krandick2024.pdf` .

[RZ] Fabrice Rouillier and Paul Zimmermann, *Efficient isolation of polynomial's real roots*, Journal of Computational and Applied Mathematics **162** (2004), 33-50.
`https://sites.math.rutgers.edu/~zeilberg/akherim/rouillier2004.pdf` .

[S] Richard P. Stanley, *"Catalan Numbers"*, Cambridge University Press, 2015.

Shalosh B. Ekhad and Doron Zeilberger, Department of Mathematics, Rutgers University (New Brunswick), Hill Center-Busch Campus, 110 Frelinghuysen Rd., Piscataway, NJ 08854-8019, USA.
Email: `ShaloshBEkhad at gmail dot com`, `DoronZeil at gmail dot com` .

Exclusively published in the Personal Journal of S.B. Ekhad and D. Zeilberger and arxiv.org.

July 16, 2024.