

THE WIDTH OF A RANDOM CLADE; OR, THE MAXIMUM
FORTUNE OF A VERY CAUTIOUS GAMBLER

By Richard F. Green

Technical Report No. 67

Department of Statistics
University of California
Riverside, California 92521

July, 1980

INTRODUCTION

In his paper, "Evolving paleontological views on deterministic and stochastic approaches," Schopf (1979) discussed the possibility that some observed paleontological patterns might be due to random events. He pointed out that if observed patterns cannot be distinguished statistically from random patterns then it is futile to seek deterministic explanations for what might actually be random patterns.

One place where pattern may be found and examined for randomness is in clades, the major branches of the evolutionary tree. Gould et al. (1977) have defined a number of clade statistics, one of which is "maximum diversity," the number of lineages existing when the clade is widest.

In order to investigate the distribution of the various clade statistics, Gould et al. (1977) simulated a number of random clades, beginning with one lineage in each simulated clade and permitting branching or extinction of lineages at random with fixed probabilities for branching and extinction. The distribution of certain clade statistics can be calculated analytically and computer simulations are not necessary. In this paper a formula is given for the distribution of the maximum diversity of a random clade, all of whose lineages become extinct after a given number of branchings and extinctions, and the distribution is calculated in several cases. The calculated distribution may be compared with the limiting distribution obtained by using a Brownian motion approximation to the underlying random walk.

THE PROBLEM

Assume that a clade starts with one lineage (is monophyletic) and that each lineage becomes extinct at random with some fixed rate and

that each lineage also branches at random with some fixed rate (which may be different from the extinction rate). Assume that all the lineages in the clade become extinct after exactly $n = 2m$ steps (branchings and extinctions) where the origin of the first lineage is counted as a step. The problem is to determine the distribution of the maximum diversity of the clade, that is, the number of lineages existing when the clade is widest.

The problem is equivalent to asking what is the maximum fortune of a gambler who plays a fair game, winning or losing one unit each play, with probability one-half each, who wins the first play and continues to play until he finally is even again, which happens after $n = 2m$ plays. Notice that if we know that after beginning play the gambler is first even again after exactly n plays, the distribution of his maximum fortune does not depend on his chances of winning or losing on each play, provided that those probabilities remain constant.

The problem is essentially a fair random walk, where a particle starts at 0 at time = 0 (or at 1 at time = 1) and takes a step right or left at random with probability = .5 each. The first step is to the right and the walk continues until the particle returns to 0. Let Y = the maximum value achieved by such a random walk. Then $P(Y < k) = 1 - 1/k$, or $P(Y = k) = 1/k(k+1)$. This result is for the case in which nothing is known about how long the random walk spends to the right of zero. If it is known that the walk returns to zero for the first time at time = $n = 2m$, then the distribution of the maximum may be calculated by using the reflection principle (Feller 1957).

METHOD

The exact calculation.

Consider a fair random walk which remains positive until the $n = 2m^{\text{th}}$ step. Let $Y =$ the maximum position of the random walk for $t \in (0, 2m)$. For any $k \geq 3$ we want to find $P(Y < k)$. (In general, the cumulative distribution function is calculated for $P(Y \leq k)$, but it is convenient in this case to use $<$ rather than \leq .)

For a random walk which remains positive until its first return to zero at step $n = 2m$ the process must have value 1 after the first step and after the $2m-1^{\text{st}}$ step. Use the notation (x, t) to indicate that the random walk process is at location x at time t , and consider the paths that lead from $(1, 1)$ to $(1, 2m-1)$. If we are interested in a particular value, k , then we must consider two boundaries: $x_1 = 0$ and $x_2 = k$. We are interested in the proportion of paths from $(1, 1)$ to $(1, 2m-1)$ avoiding $x_1 = 0$ which also avoid $x_2 = k$.

Denote the set of paths going from $(1, 1)$ to $(1, 2m-1)$ by Ω , and consider the subsets of such paths:

- B_1 : {random walk hits x_1 } ,
- B_2 : {random walk hits x_2 } ,
- $B_1 B_2$: {random walk hits x_1 , then hits x_2 } ,
- $B_2 B_1$: {random walk hits x_2 , then x_1 } ,
- $B_1 B_2 B_1$: {random walk hits x_1 , then x_2 , then x_1 } ,
- etc.

Denote the number of paths in each set by $N(\Omega)$, $N(B_1)$, $N(B_2)$, $N(B_1 B_2)$, etc.

Then we have

$$\begin{aligned}
 N(\Omega) &= \binom{2m-2}{m-1}, \text{ and, using the reflection principle:} \\
 N(B_1) &= \binom{2m-2}{m}, \\
 N(B_2) &= \binom{2m-2}{m+k-2}, \\
 N(B_1 B_2) &= \binom{2m-2}{m+k-1} = N(B_2 B_1), \\
 N(B_1 B_2 B_1) &= \binom{2m-2}{m+k}, \\
 (1) \quad N(B_2 B_1 B_2) &= \binom{2m-2}{m+2k-2}, \\
 N(B_1 B_2 B_1 B_2) &= \binom{2m-2}{m+2k-1} = N(B_2 B_1 B_2 B_1) \\
 N(B_1 B_2 B_1 B_2 B_1) &= \binom{2m-2}{m+2k}, \\
 N(B_2 B_1 B_2 B_1 B_2) &= \binom{2m-2}{m+3k-2}, \\
 N(B_1 B_2 B_1 B_2 B_1 B_2) &= \binom{2m-2}{m+3k-1} = N(B_2 B_1 B_2 B_1 B_2 B_1), \\
 &\text{etc.}
 \end{aligned}$$

We have

$$(2) \quad P(Y < k) = \frac{N(\Omega) - N(B_1) - N(B_2) + N(B_1 B_2) + N(B_2 B_1) - N(B_1 B_2 B_1) + \dots}{N(\Omega) - N(B_1)}.$$

Notice that after the term $N(\Omega)$ the numerator involves adding the number of paths that hit an even number of boundaries and subtracting the number of paths that hit an odd number of boundaries. Explicitly, we have

$$(3) \quad P(Y < k) = \frac{\binom{2m-2}{m-1} - \sum_{i=0}^{\infty} \binom{2m-2}{m+ik} - \sum_{i=1}^{\infty} \binom{2m-2}{m+ik-2} + 2 \sum_{i=1}^{\infty} \binom{2m-2}{m+ik-1}}{\binom{2m-2}{m-1} - \binom{2m-2}{m}}.$$

Notice that while the indices of summation run to ∞ only a finite number of terms will actually be included since $\binom{n}{k} = 0$ for $k > n$.

With the help of a computer formula (3) is easy to use to find the distribution of Y for various values of $n = 2m$. Computed values for the function $P(Y < k)$ are given in Table 1 for $k = 3, 4, 5, \dots$ for walks of length $n = 2m$ for $m = 10, 20, 40, 80$.

The same values are plotted in Figure 1.

A comparison of the cases $m = 10$ and $m = 40$ or of the cases $m = 20$ and $m = 80$ reveals that the distribution of Y has a very similar shape for different values of m , but that as m is multiplied by 4 the value of k for which $P(Y < k)$ has a certain value is multiplied (roughly) by 2. As the number of steps $n = 2m$ increases the random walk may be approximated by Brownian motion (see Karlin 1969 for a nice discussion of Brownian motion).

The Brownian motion approximation.

A random walk may be approximated by Brownian motion and a random walk that returns to the origin at a given time (not necessarily for the first time, however) may be approximated by the Brownian bridge, or "tied-down Brownian motion" (see Billingsley 1968).

For simplicity I use Standard Brownian motion ($\text{Var } W(t) = t$) with $W(0) = 0$ and assume that $W(1) = 0$ and $W(x) > 0$ for $x \in (0, 1)$. Consider the maximum value achieved by the process $W(t)$ for $t \in (0, 1)$. Call this maximum Z . Then for a given value $z > 0$, we can find $P(Z \leq z)$ using the reflection principle and a formula similar to (3). We find

$$(4) \quad P(Z \leq z) = 1 + 2 \sum_{n=1}^{\infty} (1 - 4n^2 z^2) \exp(-2n^2 z^2) .$$

Values of $P(Z \leq z)$ are given in Table 2 for values of z between .1 and 2.5.

Of particular interest is the value $z = 1.74726$ since $P(Z \leq 1.74726) = .95$. In other words, 95% of all Brownian bridge paths which are always positive during $t \in (0,1)$ remain below 1.74726. This value may be compared with 1.35810 which is the value such that 95% of all Brownian bridge paths remain between -1.35810 and +1.35810 for all $t \in (0,1)$. This value corresponds to the critical value of the Kolmogorov-Smirnov statistic. What this says, roughly, is that paths constrained to stay above zero tend to achieve higher values than the absolute deviation from zero of an unconstrained (except by the requirement that $W(1) = 0$) path.

The distribution of the Brownian motion approximation for the maximum value of the process (the "maximum diversity") is given in Table 2. For the Standard Brownian bridge the distribution is given as a function of z and the maximum tends to be a value around 1 or 2. To use the approximation for a random walk with the first return to the origin at the $n = 2m^{\text{th}}$ step the probabilities then correspond to $P(Y < k) \doteq P(Z \leq z)$ where $k = n^{\frac{1}{2}}z$. This is illustrated in Table 2 for $n = 100$. The approximation is quite good.

DISCUSSION

Patterns observed in paleontology (or community ecology; see, for example, Strong, Szyska and Simberloff 1979) may be tested to see whether they are actually random. The observed patterns may be compared with corresponding patterns obtained by computer simulation. In the example considered here, the maximum diversity of a clade, an explicit expression is given for the actual random distribution and simulation is not necessary.

For a random clade, all of whose lineages go extinct after exactly $n = 2m$ steps (branchings and extinctions), the critical value for the

maximum diversity (at the .05 level) is approximately $1.75n^{1/2}$. For the .01 level and the .10 level the corresponding approximate critical values will be $2.00n^{1/2}$ and $1.62n^{1/2}$, respectively. As the number of steps increases the maximum diversity of a random clade tends to increase as the square root of the number of steps. This is an intuitively appealing result.

The case considered here is for clades all of whose lineages are extinct, but the same approach may be used to find the distribution of the maximum diversity of a monophyletic clade that has a known number of existent lineages after a known number of steps.

While the mathematical result is simple and appealing there are practical problems involving sampling. It is not possible to know exactly how many steps there are since a change from one stratum to the next might involve unobserved branchings and extinctions. For example, the apparent extinction of two lineages might involve simply the two observed extinctions, one branching and three extinctions, two branchings and four extinctions, and so on. It should be possible to estimate the actual number of steps from the observations and then use the estimated number of steps to test whether the maximum diversity is larger than that expected for a random clade.

Another problem is bias due to sampling. For clades of a given duration it might be easier to observe those actually made up of more individuals (which, in turn, might well represent more lineages). Thus, sampling bias might tend to produce an unusually large number of clades with high maximum diversity. It is not clear how to handle this problem, but the direction of this bias does suggest that the observed values for maximum diversity should tend to be slightly greater than those calculated

theoretically using the random model, even if the actual clades are random. Thus detection of non-randomness could be due to sampling bias, but failure to detect deviations from randomness would happen despite sampling bias.

LITERATURE CITED

- Billingsley, P. 1968. Convergence of probability measures. John Wiley and Sons, New York.
- Feller, W. 1957. An introduction to probability theory and its applications, Vol. 1, 2nd Edition. John Wiley & Sons, New York.
- Gould, S. J., D. M. Raup, J. J. Sepkoski, Jr., T. J. M. Schopf, and D. S. Simberloff. 1977. The shape of evolution: a comparison of real and random clades. *Paleobiology*, 3:23-40.
- Karlin, S. 1969. A first course in stochastic processes. Academic Press, New York.
- Schopf, T. J. M. 1979. Evolving paleontological views on deterministic and stochastic approaches. *Paleobiology*, 5:337-352.
- Strong, D. R., L. A. Szyska, and D. S. Simberloff. 1979. Tests of community-wide character displacement against null hypotheses. *Evolution*, 33:897-913.

Table 1

Values of $P(Y < k)$, where Y is the maximum value achieved by a positive random walk which first returns to zero at time $n = 2m$, for various value of m

	$m = 10$	$m = 20$	$m = 40$	$m = 80$
$k = 3$.0002	.0000	.0000	.0000
4	.0527	.0001	.0000	.0000
5	.3285	.0137	.0000	.0000
6	.6748	.1096	.0010	.0000
7	.8914	.3181	.0140	.0000
8	.9757	.5622	.0676	.0003
9	.9967	.7619	.1805	.0035
10	.9998	.8893	.3386	.0174
11	1.0000	.9560	.5096	.0530
12		.9851	.6643	.1178
13		.9957	.7869	.2107
14		.9990	.8741	.3232
15		.9998	.9307	.4434
16		1.0000	.9644	.5603
17			.9829	.6656
18			.9923	.7548
19			.9968	.8265
20			.9988	.8813
21			.9995	.9214
22			.9998	.9467
23			1.0000	.9688
24				.9813
25				.9891
26				.9938
27				.9966
28				.9982
29				.9991
30				.9995
31				.9998
32				.9999
33				1.0000

Table 2

Continuous (Brownian bridge) approximation to the distribution of the maximum value achieved by a positive random walk that returns to zero for the first time after $n = 100$ steps. $P(Z < z)$ is the distribution for the Brownian bridge, $P(Y < k)$ is the distribution for the random walk. Notice that the values of the distribution functions correspond for $k = n^{1/2}z$.

z	$P(Z < z)$	$P(Y < k)$	$k = 10z$
.1	.0000	.0000	1
.2	.0000	.0000	2
.3	.0000	.0000	3
.4	.0000	.0000	4
.5	.0000	.0000	5
.6	.0001	.0001	6
.7	.0031	.0024	7
.8	.0216	.0194	8
.9	.0767	.0728	9
1.0	.1779	.1738	10
1.1	.3148	.3120	11
1.2	.4652	.4647	12
1.3	.6077	.6095	13
1.4	.7286	.7319	14
1.5	.8223	.8263	15
1.6	.8896	.8936	16
1.7	.9348	.9383	17
1.8	.9633	.9661	18
1.9	.9803	.9823	19
2.0	.9899	.9913	20
2.1	.9951	.9959	21
2.2	.9977	.9982	22
2.3	.9990	.9992	23
2.4	.9996	.9997	24
2.5	.9998	.9999	25

Figure 1: Values of $P(Y < k)$, $n = 2m$ steps

