Random Walk Stopping Time

I'm not all too familiar with the stochastic process side of things (I sure hope this entire answer isn't made trivial by them lol), but we can use standard combinatorics to answer this question (the linked notes on random walks in the comments does go through a similar approach but it doesn't quite touch upon the exact counting arguments explained here).

First, let's consider a random walk where we choose only from $\{-1, 1\}$ instead, and let T be the random variable that represents the number of times, or steps, we need to choose a number until we reach a positive sum, with the last number being included. What we desire is the expected value of T, or

$$E[T] = \sum_{n=1}^{\infty} n\, P(T = n).$$

This motivates us to consider a combinatorial argument where we find all sequence of the desired form with length n. We shall denote the number of sequences of length n as $L(n)$.

Thinking it through, such a sequence must necessarily end with 1 because otherwise we would reach a positive value before the last value, contradicting our assumption that the sequence was of length n. In addition, the other $n-1$ numbers at the start of the sequence must sum to 0 for a similar reason. This is in fact not possible when $n - 1$ is odd (which is the same thing as saying that n is even), so $L(2k) = 0$.

We shall now only consider odd n, so we can write $n - 1 = 2k$, or $n = 2k + 1$. This case doesn't easily disappear, but it also isn't impossible to work with. Those familiar with combinatorics (or if you're me, through computer simulation) will realize that $L(2k + 1) = C_k$, or the kth Catalan number (the combinatorial representation with Dyck words is pretty much the same except some numbers are put to it).

To turn the counting of these sequences into probabilities, we can simply just divide by 2^n, as each choice of -1 or 1 is binary. With this, we can actually calculate the expected value with a little bit of manipulation.

In order to verify that we're not making any mistakes though, we'll verify that we have a valid probability distribution by checking that the sum of all probabilities is in fact

1. We have that

$$1 = \sum_{n=1}^{\infty} P(T = n)$$

$$= \sum_{k=1}^{\infty} P(T = 2k - 1) + P(T = 2k)$$

$$= \sum_{k=0}^{\infty} P(T = 2k + 1)$$

$$= \sum_{k=0}^{\infty} \frac{C_k}{2^{2k+1}}.$$

In order to show that this truly is one, we can use the decently familiar generating function for the Catalan numbers to help us here. Generating functionology tells us that

$$c(x) = \sum_{k=0}^{\infty} C_k x^k = \frac{1 - \sqrt{1 - 4x}}{2x},$$

so we can divide by two and plug in $x = 1/4$ to verify that this is indeed a valid probability distribution and that we haven't messed up our counting.

The reason why we look to generating functions is two-fold, though, as this also gives us a very easy way to calculate the expected value. We know that

$$E[T] = \sum_{n=1}^{\infty} n \, P(T = n)$$

$$= \sum_{k=0}^{\infty} (1 + 2k) \frac{C_k}{2^{2k+1}}$$

$$= 1 + \sum_{k=0}^{\infty} \frac{k C_k}{2^{2k}}$$

$$= 1 + \sum_{k=0}^{\infty} \frac{k C_k}{4^k}$$

$$= 1 + c'(1/4),$$

where $c'(x)$ denotes the derivative of $c(x)$. Evaluating this, though, we see a very interesting result occur, that is, this diverges. In other words, we can expect an infinite number of steps until we hit a positive sum, and this gives us good motivation to believe the same for the $\{-1, 0, -1\}$ problem.

Much of the same steps remain for the more general set of values $\{-1, 0, 1\}$, except for a few things: the counting of the sequences and thus the resulting generating function.

Because 0 is included in the set, we can no longer use the same parity argument to set $L(2k) = 0$, and this does make counting quite harder. While I cannot provide a great combinatorial argument off the top of my head, one can verify with computer simulations as well as the combinatorial interpretations on Wikipedia that $L(n) = M_{n-1}$, or the $(n-1)$th Motzkin number. While there isn't a great to work with closed form for the numbers themselves, what we do have (and can work quite well with), is its generating function. We have that

$$m(x) = \sum_{k=0}^{\infty} M_k x^k = \frac{1 - x - \sqrt{1 - 2x - 3x^2}}{2x^2}.$$

Thus, the expected value is

$$
\begin{aligned}
E[T] &= \sum_{n=1}^{\infty} n\, P(T = n) \\
&= \sum_{k=0}^{\infty} (k+1) \frac{M_k}{3^{k+1}} \\
&= \frac{1}{3} + \frac{1}{3} \sum_{k=0}^{\infty} \frac{k M_k}{3^k} \\
&= \frac{1}{3} + \frac{1}{3} m'(1/3),
\end{aligned}
$$

and as expected (although slightly unfortunately), this diverges (as does any higher moments such as the variance).

From computer simulations though, we probably could have guessed as much. Assuming the distribution actually had finite mean and/or variance (even if they were large), the law of large numbers would lead us to think that with a decently large number of trials in computer simulation, the estimated average would go to the expected value, or at least, increasing the number of trials would hone in on some number. In reality, though, the simulations gave numbers bouncing all over the place, which was a good hint that something fishy was going on.