An Brief Investigation of Rare, Random Events

Jitendra Subramanyam

datatiq@gmail.com

ABSTRACT

In his book, *The Better Angels of Our Nature,* Steven Pinker poses a brainteaser about rare events. We explore this problem using simulations. The simulations reveal that the correct theoretical answer is not always the right answer. Furthermore, as the probability of an event becomes lower and lower, the theoretical answer no longer holds. Simulations of these rare, random events show the gap between theory and practice; they become essential when decisions need to be made on the basis on known probabilities given that theoretical calculations can lead one astray.

On page 202 of his book *The Better Angles of Our Nature* (2011, Penguin Books), Steven Pinker poses the following brainteaser to test our intuitions about rare, random events. "Suppose you live in a place that has a constant chance of being struck by lightning at any time throughout the year. Suppose the strikes are random: every day the chance of a strike is the same and the rate works out to be one strike a month. Your house is hit by lightning today, Monday. What is the most likely day for the ***next*** bolt of lightning to strike your house?" [The emphasis is mine.]

As Pinker shows (pp.202-3), the theoretically correct answer to this question is: the very next day! This unintuitive solution shows how difficult we find it to parse random events, especially those that are rare.

It is instructive to explore this problem of rare and random events by simulating them. When we do so, we find that the theoretically correct answer is not always the right answer. Furthermore, we can estimate how likely it is for the theoretically correct answer to be wrong. We will also discover an irony – when events are really rare, the correct answer diverges wildly from the Pinker's theoretical answer.

The Simulation

A year is taken to be 365 days. Since the probability of a strike works out roughly to be once a month, the probability of a strike on any given day can be approximated as 1/30. To simulate a year of strikes (or no strikes) we generate a sequence of ones and zeros with the probability of a 1 set at 1/30 and the probability of a zero set at 29/30 (which is 1 minus the probability of a one).

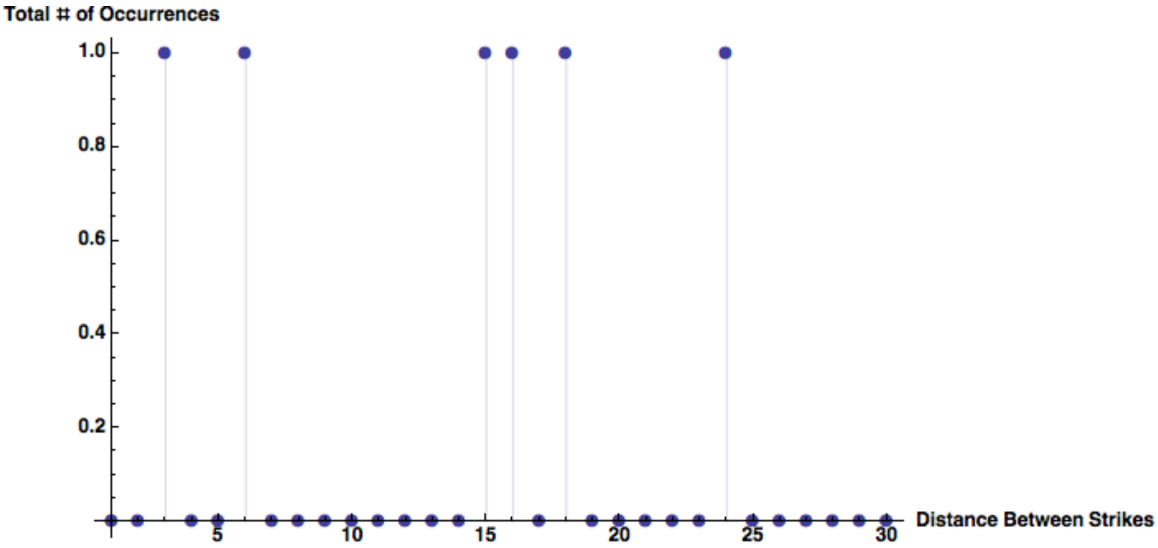
A series of such ones and zeroes will have ones followed by zeroes and then a one again. Or there might be two ones right next to each other – this is the event that the theoretical solution tells us is the most likely. To check this via the simulation we look at a series of 365 ones and zeroes and find the distances between the ones in that sequence. A striking distance of 1 means that two ones occur right next to each other – a strike today and a strike tomorrow. A distance of 2 means that a one is followed by a zero and then a one. A distance of 3 means that a one is followed by two zeros and then a one, and so on.

For each trial we calculate the number of times each of the distances occur. So a distance of one might occur 23 times, a distance of 2 might occur 19 times, and so on. We do this for distances from 1 through 20. Let's look at some simulations for a single trial. This will give us a feel for how strike distances behave.

Single Trial Simulation

For a single trial (a string of 365 ones and zeros), the number of occurrences for strike distances from 1 to 30 looks like this.

Figure 1: Single Trial, Probability of Strike = 0.03



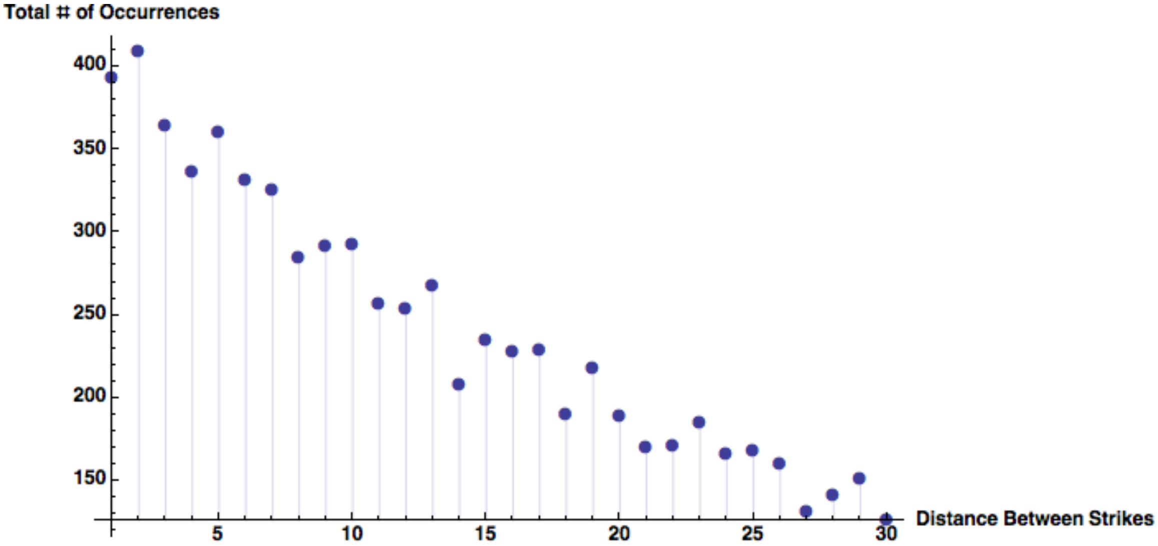
Not very interesting, but nonetheless instructive. Most strike distances including the strike distance of 1 have no occurrences at all. But one trial is only one roll of the dice, so to speak. What happens if we had 1,000 trials and then counted up the number of occurrences of strikes at distances of 1 through 30?

Multiple Trial Simulations

When we count up the number of occurrences of each strike distance across the entire set of 1,000 trials (i.e. 1,000 lists of ones and zeros, each list containing 365 ones and zeros), we get something a lot more interesting.

Notice how the number of occurrences drops off exponentially as the strike distance increases from 1 to 30. As Pinker puts it on page 203 of his book *The Better Angels of Our Nature*, "[I]n a Poisson process the intervals between events are distributed exponentially: there are lots of short intervals and fewer and fewer of them as they get longer and longer." Exactly.

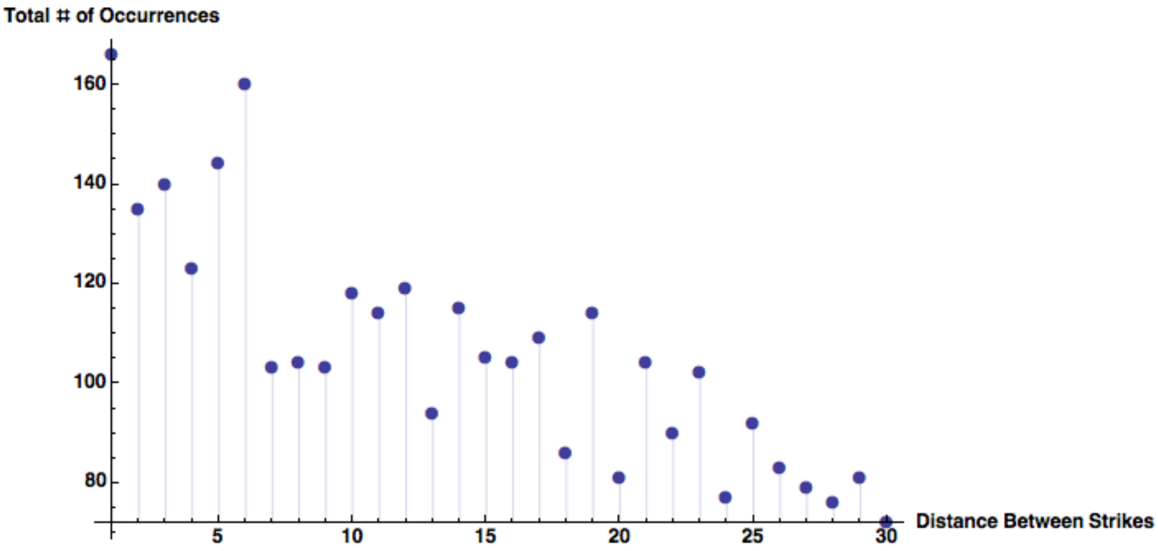
Figure 2: Multiple Trials, Probability of Strike = 0.03



Recall that the ones are quite rare in this simulation – they occur only with probability 1/30 or 0.033. For this set of 1,000 trials, a strike distance of 1 is not the most likely – it occurs 390 times versus 410 occurrences for a strike distance of 2. Strike distances 3 and 5 occur quite a bit and then the occurrences of distances greater than 5 drop off exponentially.

What happens when the strike event is still rarer? Here is what we get when the probability of a strike is 0.02.

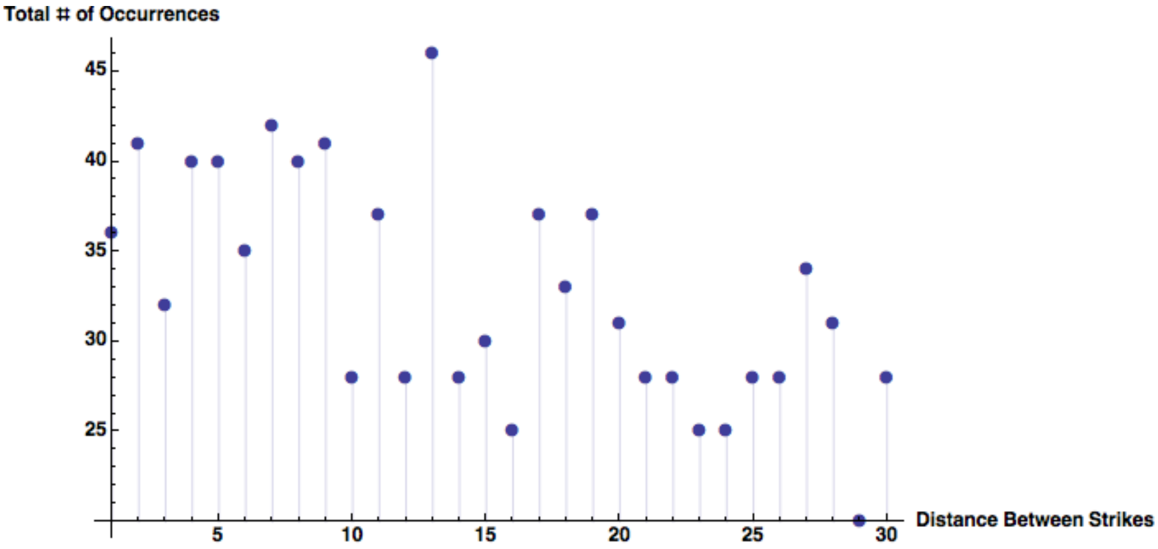
Figure 3: Multiple Trials, Probability of Strike = 0.02



The distance of 1 is now the winner. But notice that a strike distance of 6 is the second most occurring distance. Figure 3 above looks noisier than Figure 2. This is quite typical of random events and it's not clear what conclusion to draw from this other than that there is a certain amount of unpredictability in these results and were you to bet on a result it would be wise to distribute your money across the distances 1 to 5.

What happens when things get even rarer? Here is a simulation where the probability of a strike is 0.01.

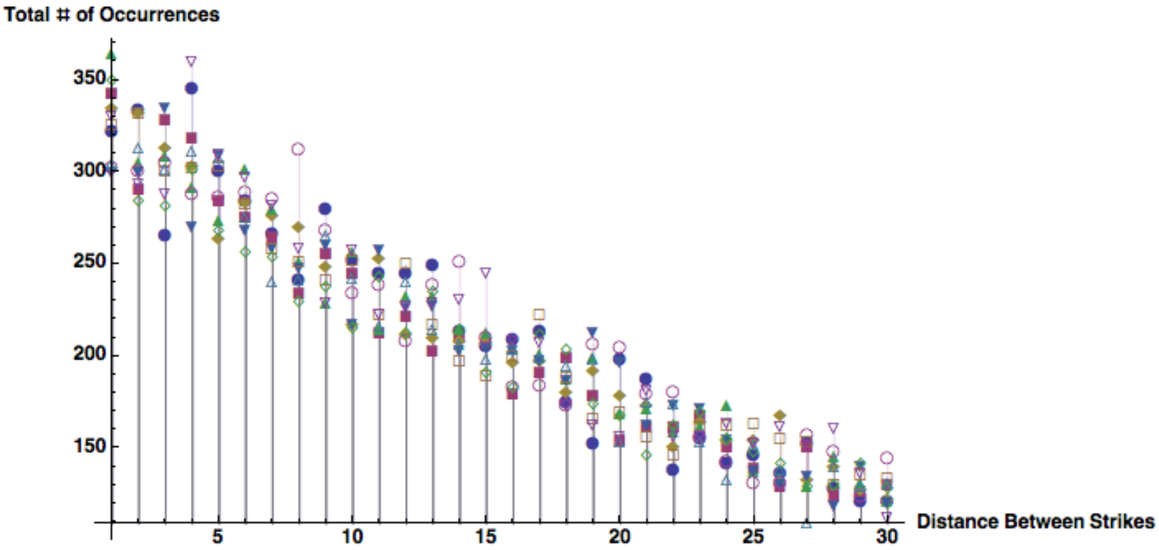
Figure 4: Multiple Trials, Probability of Strike = 0.01



Now things seem even murkier. Who would have expected distance = 13 to be the one that occurs the most? Is it the case that distance = 1 loses its importance as the probability of a strike gets lower and lower?

The way to reduce the noise is to show multiple simulations on the same plot. For a strike probability of 0.03, here is what a plot of multiple simulations looks like.

Figure 5: Ten Simulations, Probability of Strike = 0.03



Here's how to read Figure 5. Pick a number on the horizontal axis, say 5. Each dot on the vertical line is a result from a simulation – there are 10 simulations, so there will be 10 dots on each vertical line on the plot. Along the vertical line, say distance = 4, each dot tells us the number of occurrences of distance = 5 in each simulation. For distance = 5, the number of occurrences vary from about 270 to 310. Similarly, for distance = 1, the number of occurrences vary from 300 to 360 occurrences. Clearly, distance = 1 has the highest number of occurrences overall. However, note that there are 2 cases where distance = 2 has a greater number of occurrences than distance = 1 and distance =4 gives distance = 1 a run for its money.

So what should you do if this were a gamble? The theoretical answer is very clear. You pick but the simulations show that it might not be wise to depend on it entirely for deciding how to bet.

What happens when the probability of a strike event becomes even smaller? Figures 6 and 7 show what happens when the probability of strike is 0.02 and then 0.01.

Figure 6: Ten Simulations, Probability of Strike = 0.02

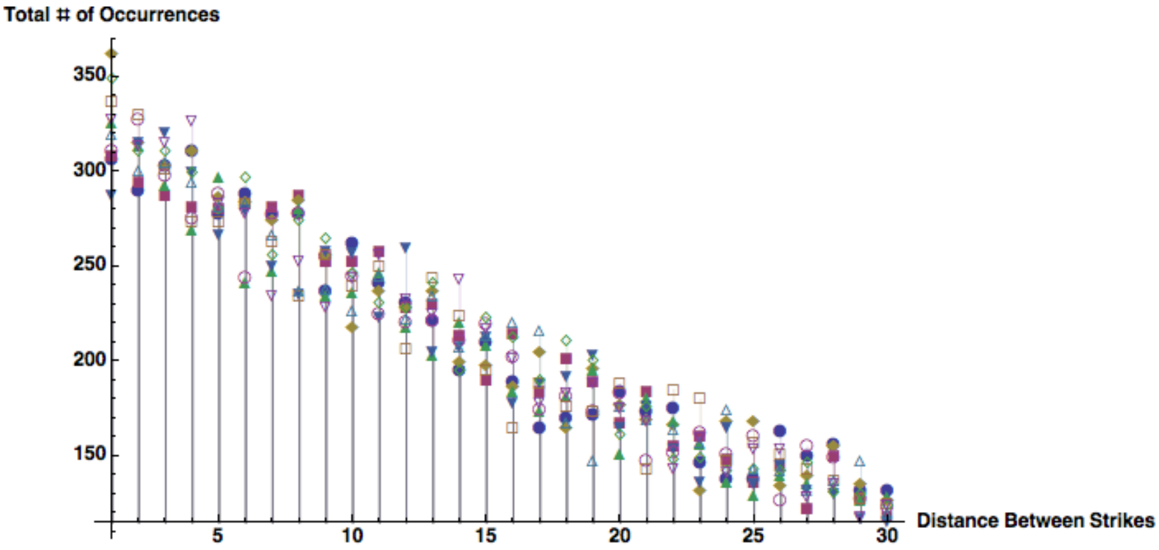
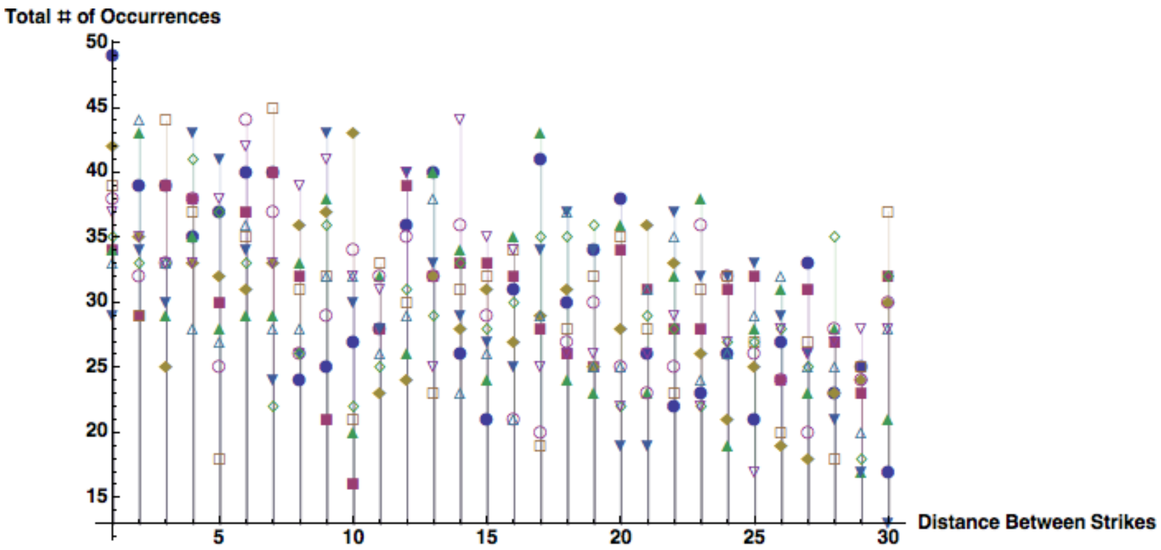


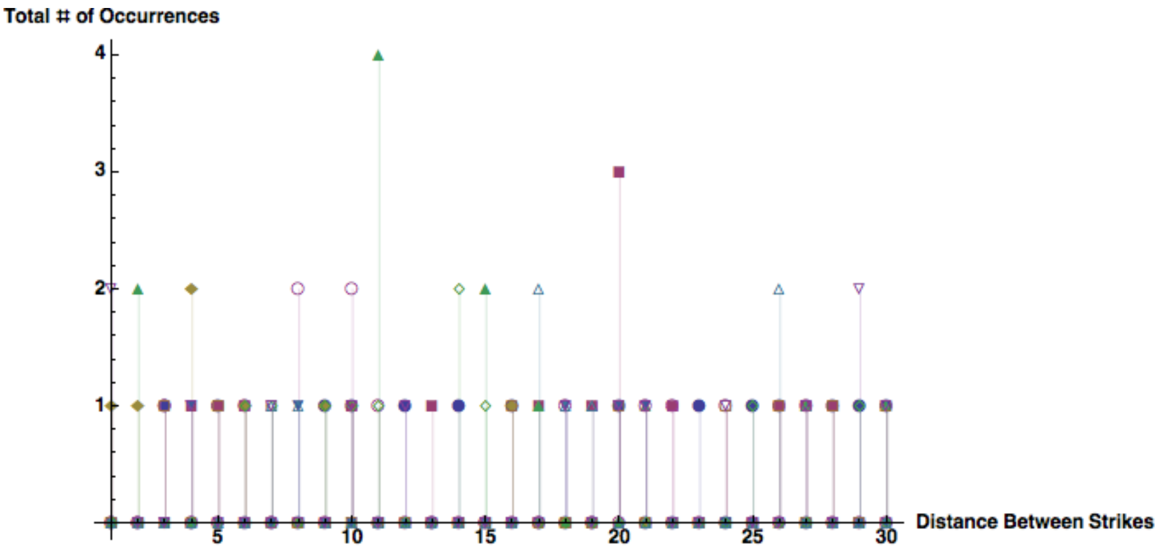
Figure 7: Ten Simulations, Probability of Strike = 0.01



The likelihood of distance = 1 becomes much murkier now. There is a set of trials for which distance = 1 has the most occurrences (around 50). But there are other sets of trials in which distances 2 through 10 (and sometimes even 14 and 17) have greater number of occurrences than distance = 1.

For an event that is rarer still, the situation becomes even more difficult to parse. Here are the results of ten simulations where the probability of a strike on any particular day is 0.001 (or about 1 strike in 1000 days).

Figure 8: Ten Simulations, Probability of Strike = 0.001

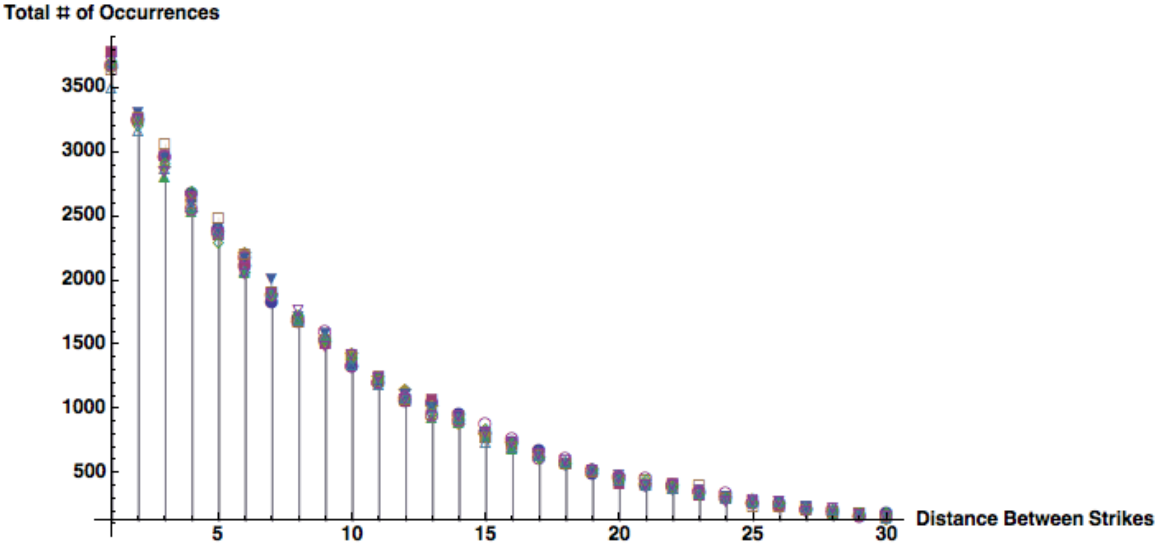


The occurrences of any of the distances between strikes are mostly zero because there are so few strikes. Further investigation may (or may not) reveal why distances of 11 and 20 spike – it might be something about this particular set of simulations; then again, it might not be.

Think of each dot as a score on a test that distance has taken. The higher the dot on the vertical line the better that distance does in the test. When the probability of a strike on any particular day is rare, how should you bet on the distance between strikes? This is a good question and requires some thought – it is not addressed in this article.

What happens when the probability of a strike increases to one strike in ten?

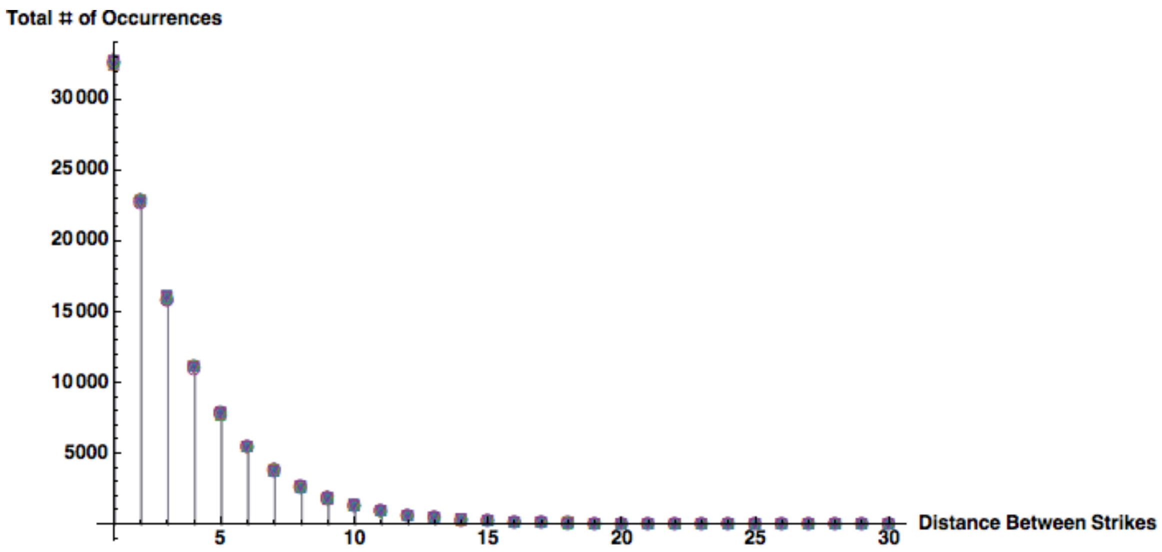
Figure 9: Ten Simulations, Probability of Strike = 0.1



A strike probability of 0.1 is about three times greater than the original strike probability of 0.03 we started with in Figure 5. What was a hint of an exponential drop in Figure 5 has taken clear shape in Figure 9. There is no longer any doubt about which strike distance to bet on: distance = 1 is the clear winner.

It is not surprising to see this pattern become more and more extreme as the probability of a strike becomes greater. Figure 10 shows the evolution.

Figure 10: Ten Simulations, Probability of Strike = 0.3



Conclusion

When events are somewhat rare, say, a 3 out of 10 chance of occurring, we can bet quite confidently that a string of two consecutive events will be more likely than any other pattern of events. As Pinker puts it, such events are well approximated by a Poisson process. However, when events get much more rare, say, a 3 out of 1,000 chance of occurring, they are no longer approximated by a Poisson distribution. Consequently, the pattern of events becomes much more complex and less predictable.