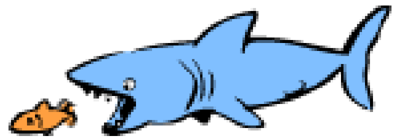
Shark attacks and the Poisson approximation

Byron Schmuland



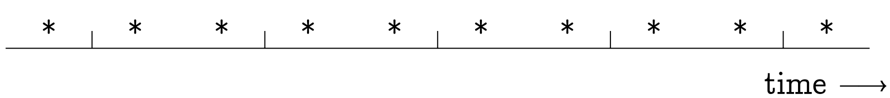
A story with the dramatic title *"Shark attacks attributed to random Poisson burst"* appeared on September 7, 2001 in the National Post newspaper. In this story, Professor David Kelton of Penn State University used a statistical model to try to explain away the surprising number of shark attacks that occurred in Florida last summer. According to this article, Kelton thinks the explanation for the spate of attacks may have nothing to do with changing currents, dwindling food supplies, the recent rise in shark-feeding tourist operations, or any other external cause.

"Just because you see events happening in a rash like this does not imply that there's some physical driver causing them to happen. It is characteristic of random processes that they exhibit this bursty behaviour," he said.

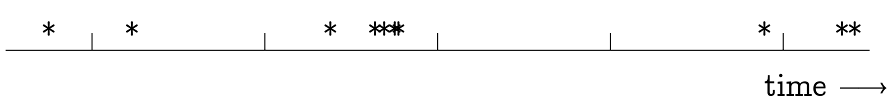
What is this professor trying to say? Can mathematics really explain the increase in shark attacks? And what are these mysterious Poisson bursts?

The main point of the Professor Kelton's comments is that unpredictable events, like shark attacks, do not occur at regular intervals as in the first diagram below, but tend to occur in clusters, as in the second diagram below. The unpredictable nature of these events means that there are bound to be periods with an above average number of events, as well as periods with a below average number events, or even no events at all.

Regular Events



Random Events



The statistical model used to study such sequences of random events gets its name from the French mathematician Siméon Denis Poisson (1781-1840), who first wrote about the *Poisson distribution* in a book on law. The Poisson distribution can be used to calculate the chance that a particular time period will exhibit an abnormally large number of events (Poisson burst), or that it will exhibit no events at all. Since Poisson's time, this distribution has been applied to many different kinds of problems such as decay of radioactive particles, ecological studies on wildlife populations, traffic flow on the internet, etc. Here is the marvellous formula that gives us information on random events:

The chance of exactly k events occurring is $\dfrac{\lambda^k}{k!} \times e^{-\lambda}$, for $k = 0, 1, 2, \ldots$

The funny looking symbol λ is the Greek letter lambda, and it stands for "the average number of events". The symbol $k! = k \times (k-1) \times \cdots \times 2 \times 1$ means the factorial of k, and $e^{-\lambda}$ is the exponential function e^x with the value $x = -\lambda$ plugged in. Let's take this new formula out for a spin.

Shark attack!

If, for example, we average two shark attacks per summer, then the chance of having six shark attacks next summer is obtained by plugging $\lambda = 2$ and $k = 6$ into the formula above; and this gives

$$\text{probability of six attacks } \approx (2^6/6!) \times e^{-2} = 0.01203,$$

which is a little over a 1% chance. This means that six shark attacks are quite unlikely in one year, though this would happen about once every fifty years. The chance that we go the whole summer without any shark attacks can also be calculated by plugging $\lambda = 2$ and $k = 0$ into the formula. This gives

$$\text{probability of no attacks } \approx (2^0/0!) \times e^{-2} = 0.13533,$$

which is a 13% chance. We expect a "sharkless summer" every seven or eight years.

In this hypothetical shark problem the number of attacks followed the Poisson distribution exactly. The Poisson distribution is most often used to find approximate probabilities in problems with n repeated trials and probability p of success. Let me show you what I mean.

Lotto 6-49

One of my favorite games to study is Lotto 6-49. Six numbers are randomly chosen from 1 to 49 and if you match all six numbers you win the jackpot. Since the number of possible ticket combinations is $\binom{49}{6} = 13{,}983{,}816$, your chance of winning the jackpot with one ticket is one out of 13,983,816, which is $p = 7.15 \times 10^{-8}$. Let's say you are a regular Lotto 6-49 player and that you buy one ticket, twice a week, for 100 years. The total number of tickets you buy is $n = 100 \times 52 \times 2 = 10{,}400$. What is the chance that you win a jackpot sometime during this 100 year run?

Now this is a pretty complex problem, but the Poisson formula makes it simple. First of all, the **average** number of jackpots during this time period is $\lambda = np = 10{,}400/13{,}983{,}816 = 0.0007437$. Plugging this in with $k = 0$ shows that the chance of a "jackpotless 100 years" is

$$\text{probability of no jackpots } \approx e^{-0.0007437} = 0.99926.$$

Wow! Even if you play Lotto 6-49 religiously for 100 years, there is a better than 99.9% chance that you never, ever win the jackpot.

Coincidences

Take two decks of cards and shuffle both of them thoroughly. Give one deck to a friend and place both your decks face down. Now, at the same time, you and your friend turn over your top card. Are they the same card? No? Then try again with the second card, the third card, etc. If you go through the whole deck what is the chance that, at some point, you and your friend turn over the same card?

In this problem there are $n = 52$ trials and the chance of a success (coincidence) on each trial is $p = 1/52$. The average number of coincidences is $\lambda = np = 52/52 = 1$, and so putting $k = 0$ in the Poisson formula gives

$$\text{probability of no coincidences } \approx e^{-1} = 0.36788.$$

The chance that you **do** see a coincidence is $1 - 0.36788 = 0.63212$. You will get a coincidence about 63% of the times you play this game. Try it and see!

Birthday problem

Suppose there are N people in your class. What are the odds that at least two people share a birthday? Imagine moving around the class checking every pair of people to see if they share a birthday. The number of trials is equal to the number of pairs of people, i.e., $n = \binom{N}{2} = N(N-1)/2$. The probability of success on a given trial is the chance that two randomly chosen people share a birthday, i.e., $p = 1/365$. This gives the average number of shared birthdays to be $\lambda = N(N-1)/2(365)$, so the probability of "no shared birthdays" is

$$\text{probability of no shared birthdays } \approx e^{-N(N-1)/2(365)},$$

and so the probability of at least one shared birthdays is approximately $1 - e^{-N(N-1)/2(365)}$. Here's what happens when you try different values of N in this expression.

Probability of a shared birthday

N	Prob	N	Prob
10	0.115991	60	0.992166
20	0.405805	70	0.998662
30	0.696320	80	0.999826
40	0.881990	90	0.999983
50	0.965131	100	0.999999

With $N = 10$ people there is only about an 11.5% chance of a shared birthday, but with $N = 30$ people there is a 69.6% chance. In a large class (like at university!) with $N = 100$ students, a shared birthday is 99.9999% certain.

For large class sizes maybe it is possible to have a triple birthday. Following the same pattern as before, we work out the chances that there is at least one triple shared birthday in a class of N people. This time you go through the class checking each triple of people, there are $\binom{N}{3} = N(N-1)(N-2)/6$ trials, and the chance of success on each trial is $p = 1/(365)^2$.

This gives $\lambda = N(N-1)(N-2)/6(365)^2$, so the probability of "no triple shared birthdays" is

$$\text{probability of no triple shared birthdays } \approx e^{-N(N-1)(N-2)/6(365)^2},$$

and so the probability of at least one triple is $1 - e^{-N(N-1)(N-2)/6(365)^2}$. Let's look at different values of N for this formula.

Probability of a triple shared birthday

N	Prob	N	Prob
10	0.000900	60	0.226522
20	0.008520	70	0.336936
30	0.030015	80	0.460278
40	0.071477	90	0.585970
50	0.136809	100	0.702915

My large first year statistics courses usually have about 100 students in them, and I always check their birthdays. According to the table, there should be a triple birthday over 70% of the time. It really is true; there is usually a triple birthday in those classes.

The Great One

During Wayne Gretzky's days as an Edmonton Oiler, he scored a remarkable 1669 points in 696 games, for a rate of $\lambda = 1669/696 = 2.39$ points per game. From the Poisson formula with $k = 0$ we estimate that the probability of Gretzky having a "pointless game" is

$$\text{probability of no points } \approx (2.39)^0/0!\, e^{-2.39} = 0.0909.$$

Over 696 games, this ought to translate to about $696 \times 0.0909 = 63.27$ pointless games. In fact, during that period he had exactly 69 games with no points.

Now for one point games we find an approximate probability of

$$\text{probability of one point } \approx (2.39)^1/1!\, e^{-2.39} = 0.2180,$$

for a predicted value of $696 \times 0.2180 = 151.71$ one point games. Let's try the same calculation for other values of k, and compare what the Poisson formula predicts to what actually happened.

Points	*Actual # Games*	*# Predicted by Poisson*
0	69	63.27
1	155	151.71
2	171	181.90
3	143	145.40
4	79	87.17
5	57	41.81
6	14	16.71
7	6	5.72
8	2	1.72
9	0	0.46

As you see there is a remarkable agreement between the predictions based on the Poisson formula, and the actual number of games with different point totals. This shows that Gretzky was not only a high scoring player, but a consistent one as well. The occasional pointless game, or occasional "Poisson burst" in seven or eight point games were not due to inconsistent play, but are exactly what is expected in any random sequence of events. Another reason why he really was the great one!