not share a birthday with any preceding people, **given** that none of the preceding people share a birthday. (There are also lots of other ways of calculating the probability.)

The probability that person 1 does not share a birthday with any preceding person is 1 since there are no preceding people. The probability of person 2 not sharing a birthday with person 1 is 364/365 (since there are 364 days that are not person 1's birthday). The probability of person 3 not sharing a birthday with person 1 or 2, given that 1 and 2 do not share a birthday is 363/365. And so on. Using X_i to represent the date of person i's birthday, we can write the probability as follows:

$$\begin{aligned}\mathbb{P}(X_1 \neq X_2 \neq \ldots \neq X_{18}) &= 1 \\ &\quad \times \mathbb{P}(X_2 \neq X_1) \\ &\quad \times \mathbb{P}(X_3 \neq X_2 \neq X_1 | X_2 \neq X_1) \\ &\quad \vdots \\ &\quad \times \mathbb{P}(X_{18} \neq X_{17} \neq \ldots \neq X_2 \neq X_1 | X_{17} \neq \ldots \neq X_2 \neq X_1)\end{aligned}$$

Putting in the relevant probabilities, we have

$$\begin{aligned}\mathbb{P}(X_1 \neq X_2 \neq \ldots \neq X_{18}) &= 1 \\ &\quad \times \frac{364}{365} \\ &\quad \times \frac{363}{365} \\ &\quad \vdots \\ &\quad \times \frac{(364-17)}{365} \\ &\approx 0.653\end{aligned}$$

So the probability that at least two people in the class share the same birthday is 1-0.653=34.7%.

4.2 Paul the octopus again

We said above that if you have 2,000 animals randomly and independently predicting the winners of 7 games, with probability 0.5 of getting the result right for each game, then the probability of at least one of them getting all 7 predictions right was 99.99998%. Where does this number come from? Well if we call the event "7 correct predictions" a "success" then the probability of a success is $p = 0.5^7$. And if we have 2,000 animals predicting independently of each other, then we have 2,000 independent trials. If we let X be the number of successes from these trials, then X has a binomial distribution with parameters $N = 2,000$ and $p = 0.5^7$ and the probability of at least one animal making 7 correct predictions is the probability of the event $X \geq 1$, i.e. $1 - F(0; N = 2,000, p = 0.5^7)$, which turns out to be 0.9999998. (Here is an `R` command that does the calculation: `1-pbinom(0,size=2000,prob=0.5^7`.)

4.3 The national lottery

It is extremely unlikely that any particular person (you, for instance) will win the Lotto lottery if you play it: each ticket has a 1 in 13,983,816 chance of winning. And yet most

weeks someone does win. So how does such an unlikely event lead to someone (who has a 1 in 13,983,816 chance of winning) winning most weeks?

About 31 million Lotto tickets are sold each week for the Saturday draw. Assuming that the choice of numbers on each ticket is independent of those on other tickets, what is the probability of there being at least one winner? We will look at two ways of calculating this.

4.3.1 Probability of a winner, using the binomial pdf

We can think of the problem as comprising $N = 31,000,000$ independent trials, in which the probability of success for each trial is $p = 1/13,983,816$. If we let X be the number of successes then X has a binomial distribution and the question we are asking is "What is $\mathbb{P}(X > 0)$?".

$$\begin{aligned}
\mathbb{P}(X > 0) &= 1 - \mathbb{P}(X = 0) \\
&= 1 - f(0; N = 31,000,000; p = 1/13,983,816) \\
&= 1 - \binom{31,000,000}{0} \left(\frac{1}{13,983,816}\right)^0 \left(1 - \frac{1}{13,983,816}\right)^{31,000,000-0} \\
&= 1 - 1 \times 1 \times \left(1 - \frac{1}{13,983,816}\right)^{31,000,000-0} \\
&\approx 0.891
\end{aligned}$$

(Here's an `R` command to do the above calculation: `1-dbinom(0,31e6,1/13983816)`.) So according to our calculation, there should be at least one winner on about 9 out of 10 weeks on average. This agrees fairly well with the data. In 2011, for example, there were 28 Saturday Lotto draws and 3 rollovers (source: http://lottery.merseyworld.com). So there was at least one winner 25/28=89.3% of the time. This is pretty close to our prediction of 89.1%, so on this evidence it would appear that our binomial model for the Lotto process is not a bad one.

4.3.2 Probability of a winner, using e

There is a quick and neat way of calculating the approximate probability using the following result (which we will not prove and which turns out to be useful later too):

$$\lim_{y\to\infty} \left(1 - \frac{1}{y}\right)^{ay} = e^{-a} \tag{1}$$

It goes as follows: The probability of any Lotto ticket being a winner is $p = 1/13,983,816$. So the probability of it not being a winner is $(1 - p)$, and the probability of there being no winners ($X = 0$) when 31,000,000 tickets have been sold is $\mathbb{P}(X = 0) = (1-p)^{31,000,000}$. If we let $y = 13,983,816$ then $\mathbb{P}(X = 0) = (1-1/y)^{31,000,000}$, which we can write as $\mathbb{P}(X = 0) = (1-1/y)^{ay}$, where $a = 31,000,000/13,983,816$. Now if y is very large then

$$\begin{aligned}
\mathbb{P}(X > 0) &= 1 - \mathbb{P}(X = 0) \\
&= 1 - (1 - 1/y)^{ay} \\
&\approx 1 - \lim_{y \to \infty} \left(1 - \frac{1}{y}\right)^{ay} \\
&\approx 1 - e^{-a}
\end{aligned}$$

Evaluating $1 - e^{-a} = 1 - e^{-31{,}000{,}000/13{,}983{,}816}$ we find that it is 0.891 (to three decimal places) – equal to the probability we calculated using the binomial distribution.

4.4 Conclusion

Although the title of this section sounds counter-intuitive, it is not really at all surprising that unlikely events happen all the time. Even though specific events (you winning the lottery, or Paul correctly predicting the outcome of 7 football games, for example) are extremely unlikely, when there are many opportunities for the event to occur (e.g. 31 million tickets, or 20 predicting animals) it becomes likely to occur.

If you ignore the fact that there were many opportunities you will draw misleading conclusions (and, for example, conclude that Paul the octopus is psychic).

5 How long must you wait for something to happen?

OK, so you know that you have a very low chance of winning the lottery in any given week, but you also know that the more opportunities you have to win, the greater the chance that you will win one at least one opportunity. So one reasonable question to ask is "What are my chances of winning if I play N times?" and you can use the binomial distribution to answer this question. But another question you might want to ask is "What is the probability that I wait no longer than k weeks before winning the lottery?"

5.1 The geometric distribution

We answer this question by first asking "What is the probability that I wait exactly k weeks until winning?" (i.e., win first on week k). This is a relatively easy question to answer since this is just the probability that you have $k - 1$ failures to win and then one success. If the probability of a success is p, and we let X be the number of failures before the first success, we can write this as $\mathbb{P}(X = k) = (1 - p)^{k-1}p$.

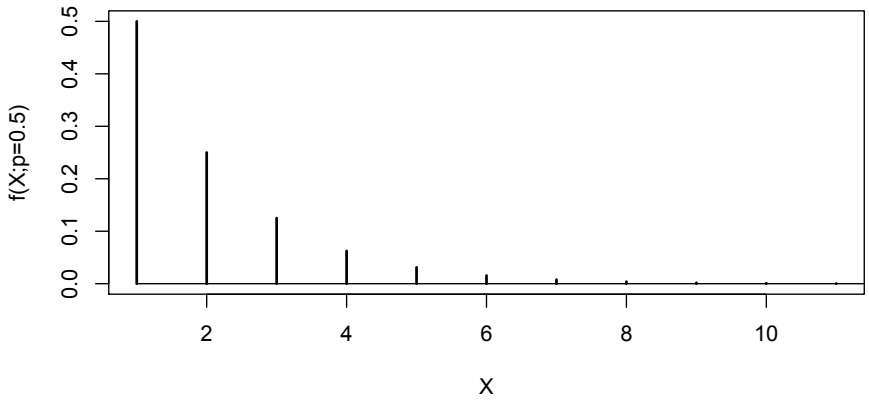
The geometric probability mass function[3]:

$$\begin{aligned}
f(k; p) = \mathbb{P}(X = k) &= (1 - p)^{k-1}p \quad \text{for } k \in \{1, 2, \ldots\} \\
&= 0 \quad \text{otherwise.}
\end{aligned}$$

[3]There are two ways of parameterising the geometric distribution. One deals with the number of trials **before** the first success; the other with the number of trials **until** the first success (which is what we do here). This Wikipedia page explains the difference: http://en.wikipedia.org/wiki/Geometric_distribution

We say that X "has a geometric distribution" or "is geometrically distributed". Figure 4 shows the geometric probability mass function (pmf) for $p = 0.5$.

Figure 4: The geometric probability mass function for $p = 0.5$.



Now on to the question we asked initially: "What is the probability that I wait no longer than k weeks until I win the lottery?". This is

$$\begin{aligned}
\mathbb{P}(X \leq k) &= \sum_{i=1}^{k} f(i;p) \\
&= \sum_{i=1}^{k} (1-p)^{i-1} p \\
&= \sum_{i=0}^{k-1} (1-p)^{i} p \\
&= \left[\frac{1-(1-p)^k}{1-(1-p)}\right] p \qquad \left(\text{since } \sum_{i=0}^{k-1} r^i = \frac{1-r^{(k-1)+1}}{1-r}\right) \\
&= \left[\frac{1-(1-p)^k}{p}\right] p \\
&= 1-(1-p)^k
\end{aligned}$$

Hence:

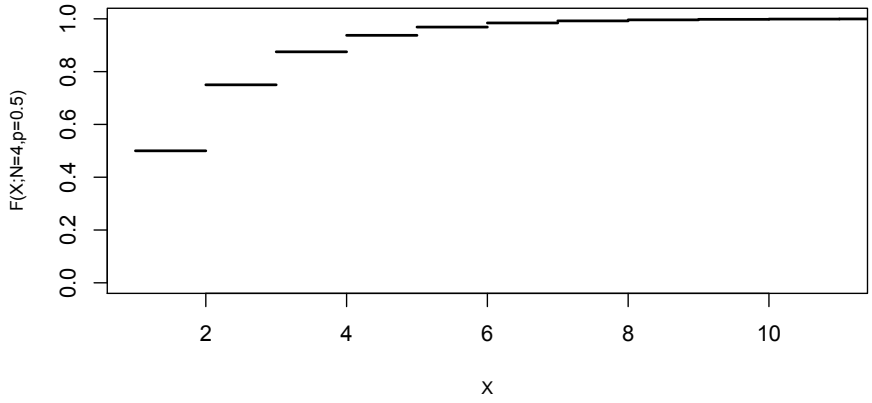
The geometric cumulative distribution function:

$$F(k;p) \;=\; \mathbb{P}(X \leq k) \;=\; \sum_{j=1}^{\lfloor k \rfloor} f(j;p) \;=\; 1-(1-p)^{\lfloor k \rfloor} \qquad \text{for } k \geq 1.$$

Figure 5 shows the geometric probability mass function for $p = 0.5$.

Here's another way of thinking about the geometric cdf. The question "What is the probability that I wait no longer than k weeks until I win the lottery?" is the same question as "What is the probability that not all of the first k trials are failures?" And using the binomial distribution we have $\mathbb{P}_{binom}(X \leq k) = 1 - \binom{k+1}{0} p^0 (1-p)^k = 1-(1-p)^k$.

Figure 5: The geometric cumulative distribution function for $p = 0.5$.



Lets calculate this probability for that you wait no more than two years at 52 lottery draws a year (so $k = 104$ weeks): $F(100, p = 1/13,983,816) = 1 - (1 - 1/13,983,816)^{100} \approx 0.000007$, i.e., about 7 hundred-thoustandths of a percent, or 7 in a million.

6 An unusual number of murders in London?

After four murders occurred in a single day in London, a 2008, a BBC News report said "Last year the Metropolitan Police recorded 160 homicides - about three every week. To have four fatal stabbings in one day could be a statistical freak, said BBC correspondent Andy Tighe." (source: http://news.bbc.co.uk/1/hi/uk/7502569.stm).

When you have an average of about three murders a week, is four in a day a very unlikely event, or is Andy Tighe making more of it than he should? Lets look at the data and the associated probabilities. We can get the date of every murder in London from January 2006 for five years from this Guardian website. I downloaded the data and show in Figure 6 the cumulative number of murders plotted against days since 1st January 2006, for 1,000 days.

The fact that the line in Figure 6 has a slope of 0.458 bears out the Metropolitan Police's statement that murders occur at a frequency of about three a week on average. This is because the slope is the total number of murders divided by total number of days and hence implies that there are on average 0.458 murders a day, or put another way, about one murder every $1/0.458 \approx 2.2$ days , or three murders every 6.6 days (which is close to a week) on average.

So if murders are occurring at an average rate of 0.458 a day, is it very unlikely that four occur in one day? To answer this question we need to build a statistical model for the random variable X="number of murders in a day". This presents us with a problem – because the only statistical models we have developed so far are for binary ('success"/"failure") outcomes (e.g. Paul predicts the result of a game correctly or not; you win the lottery or not). The binomial distribution can deal with an outcome that is a count (e.g. number of times Paul predicts correctly, or number of lottery wins), but only when this number is made up from binary outcomes from a number of discrete units (games for Paul, or weeks for the lottery).