The strong birthday problem

This issue marks the tenth birthday of *Significance* magazine. To mark it, Mario Cortina Borja examines the strong birthday problem.

The birthday problem

The classical birthday problem is well known among statisticians and even among some school children. It often comes as one of their early lessons in probability - perhaps because it involves birthdays, which all children like, and it has a somewhat unexpected answer. It is also fairly simple to solve. It can be expressed as follows: "How large a group of people do you need to make it more likely than not that two of them share a birthday?" Or, in a classroom: "How big does this class have to be before two children are likely to have the same birthday?" Most children (and most adults) would guess the answer to be around half of 365 - call it 183, since you cannot have half-people; of course they are way out. The correct answer is 23 - see the box. In a class of 23 children, it is more likely than not that two of them have the same birthday. On a football pitch there are 23 players if you count the referee; in half of all football matches there will be at least one shared birthday on the pitch.

So far so simple. And we can continue beyond 50-50 chances: in a party of 57 people there is a 0.99 probability of a shared birthday; 70 people are enough to give a 0.999 probability. And of course if you have 365 (or 366 if you really want leap years) people in a room, a shared birthday is certain. Even for a group of 100 people the probability is 0.999999693, which is as close to one as makes no real difference.

However, we have made two assumptions in all of this. Firstly, we supposed that birthdays are independent - that the date of one birthday does not affect the date of another. This would not be true if we had done the experiment at a convention of twins and triplets - one twin's birthday would determine the birthday of someone else in the room. We have also assumed equiprobability: that birthdays are evenly distributed throughout the year - that January 1st, April 23rd and September 19th are all equally likely as birthdays, each with a probability of 1/365.

The assumption of uniform birth dates simplifies these calculations considerably; however, it is not actually true. There is much evidence that it does not hold for real human populations; birthday distributions actually depend on social, religious, economic and environmental factors. Figure 1 illustrates these patterns for...
children born in 2011 in four European and four American countries. It uses the adjusted proportions of months of birth and approximate month of conception, and takes into account the different number of days in months.

The countries shown have substantial variations in the months when peaks occur, in their amplitudes, and in the deviations from the hypothesis of uniformity. Western Europe has birth-peaks from July to October, with spring troughs; in the Americas, Brazil’s births are distributed in an opposite pattern. These differences may reflect environmental factors, notably temperature, which has an effect on human fertility and on human characteristics (e.g., males sharing their birthday with females in a population). Two excellent reviews of such generalisations and methods to study them are available. In this article we review what is known as the “strong birthday” problem as seemingly first defined by Anirban DasGupta in 2005. It refers to the probability that not just one person but everybody in a group of \( n \) individuals has a birthday shared by someone else in the group. Nobody in the group has a lone birthday. Clearly you need more people for this than for the classical birthday problem: \( n \) must be greater than 23, but how much greater?

Suppose that there are \( m \) classes (for example, \( m = 365 \) for daily dates of birth), in which a group of \( n \) people fall with equal probability \( p_r = 1/m \) for \( r = 1, \ldots, m \). Let \( N \) be a random variable which counts the number of unique individuals (those whose birthdays are not shared by anybody else) out of a group of size \( n \); \( N \) must lie between 0 and whichever of \( m \) or \( n \) is smaller. It can be shown that

\[
\Pr(N = k) = \sum_{i=0}^{\min(k, n)} \frac{(-1)^{k-i} \cdot i!}{k!(i-k)!} \frac{m!n!(m-i)^{-i}}{i!(n-i)!m^n}
\]

and using this expression for \( N = 0 \), the probabilities of everybody in a group of \( n \) individuals having a shared birthday are given in Table 1.

The strong birthday problem

There are many generalisations and variants of the birthday problem. For example, one can study multiple coincidences (for example, more than two people having the same birthday: if you want three children to be likely to share a birthday you need a class of 88), or almost-coincidences (the number of people required to have at least two persons whose birthdays fall at most one or two or three days apart), or birthdays shared among people of different sizes. Clearly you need more people for this than for the classical birthday problem: \( n \) must be greater than 23, but how much greater?

**Figure 1.** Adjusted proportions of month of birth and approximate month of conception in (a) four European and (b) four American countries, 2011. Note that these are monthly cross-sectional data and these national patterns may vary across short periods of time. Data from the UN Statistics Division, available at [http://data.un.org/Data.aspx?d=POP&f=tableCode:55](http://data.un.org/Data.aspx?d=POP&f=tableCode:55), accessed November 2013.
Note that if \( n = 1 \) the probability of a lone birthday is trivially 0; for \( n = 2 \) it is \( 1/365 = 0.0027; \) thereafter it decreases, reaching its minimum around \( n = 365 \), and rising thereafter—first at a very slow rate, quickly gathering such as a there are actually some unfeasible in fact, Those interested can Two of the players will probably share a birthday. Hieu Le/Stock/Thinkstock.

<table>
<thead>
<tr>
<th>( n )</th>
<th>Probability</th>
<th>( n )</th>
<th>Probability</th>
<th>( n )</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>( 2 \times 10^{-10} )</td>
<td>3200</td>
<td>0.608555</td>
<td>5200</td>
<td>0.996683</td>
</tr>
<tr>
<td>2000</td>
<td>0.000152</td>
<td>3400</td>
<td>0.737717</td>
<td>5400</td>
<td>0.998009</td>
</tr>
<tr>
<td>2200</td>
<td>0.004134</td>
<td>3600</td>
<td>0.830403</td>
<td>5600</td>
<td>0.998805</td>
</tr>
<tr>
<td>2400</td>
<td>0.032837</td>
<td>3800</td>
<td>0.892940</td>
<td>5800</td>
<td>0.999286</td>
</tr>
<tr>
<td>2600</td>
<td>0.120095</td>
<td>4000</td>
<td>0.933491</td>
<td>6000</td>
<td>0.999730</td>
</tr>
<tr>
<td>2800</td>
<td>0.269630</td>
<td>4200</td>
<td>0.959126</td>
<td>7000</td>
<td>0.999968</td>
</tr>
<tr>
<td>3000</td>
<td>0.445765</td>
<td>4400</td>
<td>0.975064</td>
<td>8000</td>
<td>0.999998</td>
</tr>
<tr>
<td>3063</td>
<td>0.499923</td>
<td>4600</td>
<td>0.984867</td>
<td>9000</td>
<td>1.000000</td>
</tr>
<tr>
<td>3064</td>
<td>0.500766</td>
<td>4800</td>
<td>0.990851</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3065</td>
<td>0.501607</td>
<td>5000</td>
<td>0.994484</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Probabilities of no lone birthdays for the strong birthday problem

The strong birthday problem with unequal probabilities

The strong birthday problem for no lone birthdays with an unequal probability distribution of birthdays is very hard indeed. The following expression is available, where \( m \) is the number of classes (again 365):

\[
\Pr(N = 0) = 1 - \sum_{j=1}^{\min(m, n)} \left(-1\right)^{j-1} \frac{n!}{j!(n-j)!} \prod_{a \neq j} \left(1 - p_a - p_j - \cdots - p_j\right)
\]

Though the mathematics is straightforward, the computational cost of solving it is fiendish - unfeasible in fact. Those interested can see it investigated on the Significance website at www.significancemagazine.com.

However, as in the classical birthday problem, an unequal distribution of birthdays makes only a small difference; using a uniform distribution of birthdays approximates well to most real-life cases, unless the birthdays are concentrated on a few high-frequency days with the other days occurring very rarely. (It would not work for a gathering of astrologers born under Aries, for example.) Using equation (1) I calculated the probability of no lone birthdays for the real-life birthday distributions shown in Figure 1; as well as for several other empirical distributions. I did it for gatherings ranging from 1 to 5000 people and compared these probabilities to those from the uniform case; all along the range the absolute differences were less than 1%. Besides this empirical evidence, it is possible to develop a formal argument to show that the adjustments required to cope with unequal probabilities are as small as those in the classical birthday problem.

Conclusions

The answer to the strong birthday problem is 3064. This is the number of people that need to be gathered together before there is a 50% chance that everyone in the gathering shares his birthday with at least one other. The non-uniform spread of birthdays through the year has very little effect on the strong birthday problem; so it seems that 3064 is very close to reality for virtually all populations of interest.

Apart from the obvious areas in which this solution might be seen as useful — such as a particularly brilliant answer in the University Challenge TV quiz, or as a fascinating topic of pub conversation — there are actually some potentially important applications. DasGupta has mentioned using the strong birthday problem methodology in criminology, for instance in calculating the expected number of matched physical features in looksalikes. Another application is in record linkage problems, where the distribution of the number of non-matched individuals with data in two files is of interest. This can be used in cryptographic attacks on digital signatures, for example. Both lie beyond the scope of this brief review; but happy birthdays do not.

References


Two of the players will probably share a birthday. Hieu Le/Stock/Thinkstock.

Mario Cortina Borja is Senior Lecturer in Statistics at the Institute of Child Health, University College London.