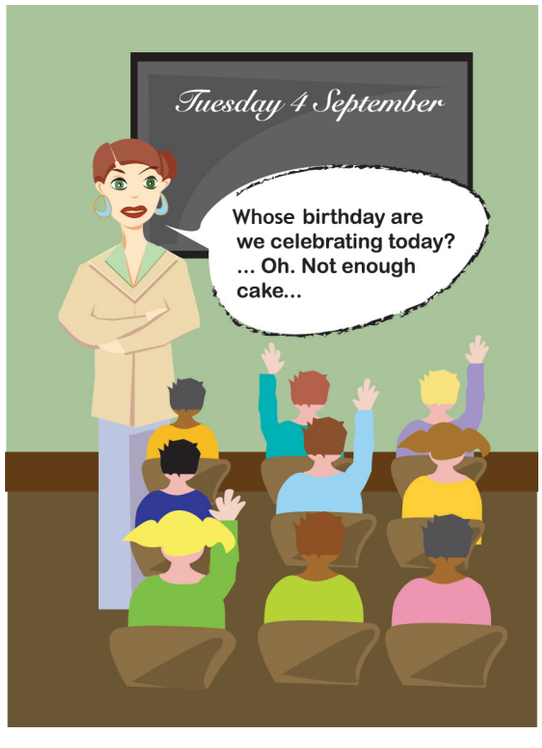
The birthday problem

Happy Birthday! If it is not your birthday today, it may well be the birthday of someone in your office, or in your class at school. Possibly, it is the birthday of *two* people in your office or your class. Just how big does a group have to be before its members are likely to share a birthday? **Mario Cortina Borja** and **John Haigh** explain the birthday problem.

When students first encounter the question "How large a group of people do you need to make it more likely than not that two of them share a birthday?", most answers offered are far too big. A fairly common suggestion is 366/2 = 183, and the normal reaction to the assertion that the answer is 23 is disbelief.

One way to reach this answer is given in the box. Statisticians who prefer more technical notation might prefer this version of it:

One way to reach this answer is given in the box. Statisticians who prefer more technical notation might prefer the following well-known version of it. Given N equiprobable cells, the chance that r balls thrown randomly into them will all fall in *different* cells is $D(N,r) = N(N-1)(N-2)\ldots(N-r+1)/N^r$; hence, the chance that at least two will fall in the same cell is the complement of this—and for $N = 365$ or $N = 366$, this complement first exceeds 50% when $r = 23$. An intuitive explanation as to why this should be no real surprise is that there are as many as 253 *pairs* among 23 balls – balls 1 and 2, 1 and 3, 1 and 4, … 1 and 23, then balls 2 and 3, 2 and 4, … 2 and 23, then balls 3 and 4, 3 and 5 and so on. Each pair has chance 1/365 of falling



Cartoon: Caroline Hilde

Solving the birthday problem

Suppose a class of children announce their birthdays one by one. The first child reveals his special day. The second child reveals his. The probability of them being the same is 1/365. The probability of them being different must therefore be the converse of that, 364/365.

When two different birthdays have been announced, the third child speaks. The probability of his being different from either of the others is 363/365, as there are now only 363 different days left for him. So the probability of all three birthdays being different is 364/365 x 363/365.

The probability of the fourth birthday also being different is the previous answer multiplied by 362/365; and so on down the rest of the class. By the time the 23rd boy gives his birthday, the odds that all of them so far are different are 364/365 x 363/365 x 362/365 x 361/365 …. x 343/365. This comes out to just under a half. In other words, there is a less than 50-50 chance that all the birthdays so far are different.

And that in turn means that there must be a greater than 50-50 chance that the opposite is true, and that there is a shared birthday somewhere among those 23 children.

Table 1. For a range of values of p, r is the smallest number of balls to be thrown randomly among N = 365 cells so that the chance that some cell contains more than one ball is at least p

p	0.500	0.750	0.900	0.950	0.990	0.999
r	23	32	41	47	57	70

in the same cell—together, these 253 chances make it more likely than not that at least one of them will happen.

The same approach, of course, leads to the values shown in Table 1: how many people are needed so that the chance that two of them share a birthday is at least p?

Empirical birthday distributions

The assumption, when applying these calculations to birthdays, is that all possible birth dates are equally likely. But this assumption is false, as shown by diverse data taken from many countries and at different times—for an example see Figure 1, showing monthly births in England and Wales in 1979 and 2005. (Our data are from the Office for National Statistics. The figures were adjusted to take account of the different numbers of days in each month and are the normalised proportions of yearly births falling in each calendar month.)

The overall picture illustrates the change in England and Wales from the "European" pattern (more births in spring and early summer with a secondary peak in September) in 1979 to the "American" pattern (the peak in September being dominant) in 2005. From 2001 to 2005, the day of the year with most births was consistently between 23rd and 27th September.

These aggregated monthly frequencies hide two further sources of variation—holidays and day of the week. The holiday effect operates on two levels: the rise in birth numbers in late spring and again in late September points to conceptions around the summer and Christmas holidays (in his *Untold Stories*, Alan Bennett writes "[my birthday]... is on May 9 and my brother's too, though he is three years older than I am. The coincidence is always good for a laugh, particularly when it dawns that we must both have been conceived during the old August Bank Holiday, sex confined to the holidays perhaps..."); second, there are fewer births on public holidays owing to a reduction in hospitals' activity—Boxing Day had the lowest number of births in 25 of the 27 years from 1979 to 2005. Figure 2 covers the 17 517 192 births that took place in England and Wales between 1979 and 2005 (excluding the 12 311 births that occurred on February 29th).

> The rise in births in late September points to conception around the Christmas holidays

The variation between days of the week is shown by the boxplots in Figure 3: more births than average occur between Tuesday and Friday and fewest on weekends. But because a year is not an exact number of weeks, the amalgamated data over many years hide this weekly cycle.

Another difference between 1979 and 2005 is in the variability in the proportions of births occurring on a single day. Taking the most extreme days in each year, 1979 saw around 4.05% of births on the 20 days with fewest births, and 6.42% on the 20 with most; the figures for 2005 are 4.40% and 6.28%, respectively.

Adjustment for non-uniform distributions

Given this clear evidence that the simple assumption of a uniform distribution of birth dates across the year is wrong, how does this affect the answer to the initial question posed? There are many different ways of showing that any departure from uniformity will tend to *increase* the chance that a group of size r contains a pair with a common birthday[1]. Hence, because 23 suffices when there are 366 equally likely dates, it certainly suffices when some of those dates are more likely than others.

Nunnikhoven[2] examined how well the simple model approximates reality. If we write the true frequency that any ball falls in cell j as $p_j = 1/N + d_j$, where $\sum d_j = 0$, and define $S = \sum d_j^2$ as a measure of the variability of the daily frequencies, then, to first-order approximation, the chance that all r balls fall in different cells is

$$D(N,r)\{1-\frac{r(r-1)SN}{2(N-1)}\}$$

For the amalgamated data from 1979 to 2005, $S = 3.59 \times 10^{-6}$; since $D(365, 22) = 0.523$, this extra consideration makes hardly any difference at all—

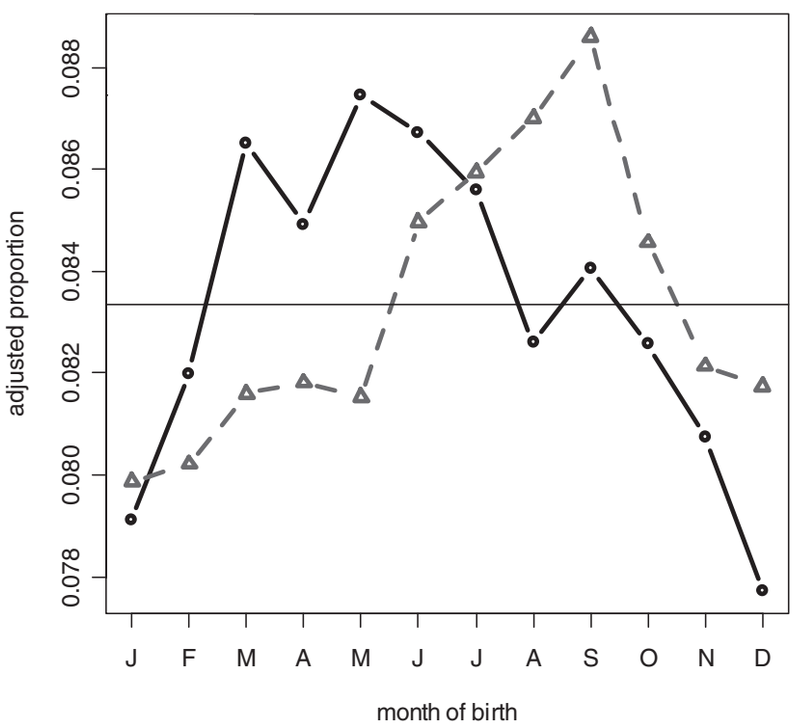


Figure 1. Adjusted monthly proportions of births: England and Wales, 1979 (●)and 2005 (▲); the horizontal line corresponds to a uniform distribution

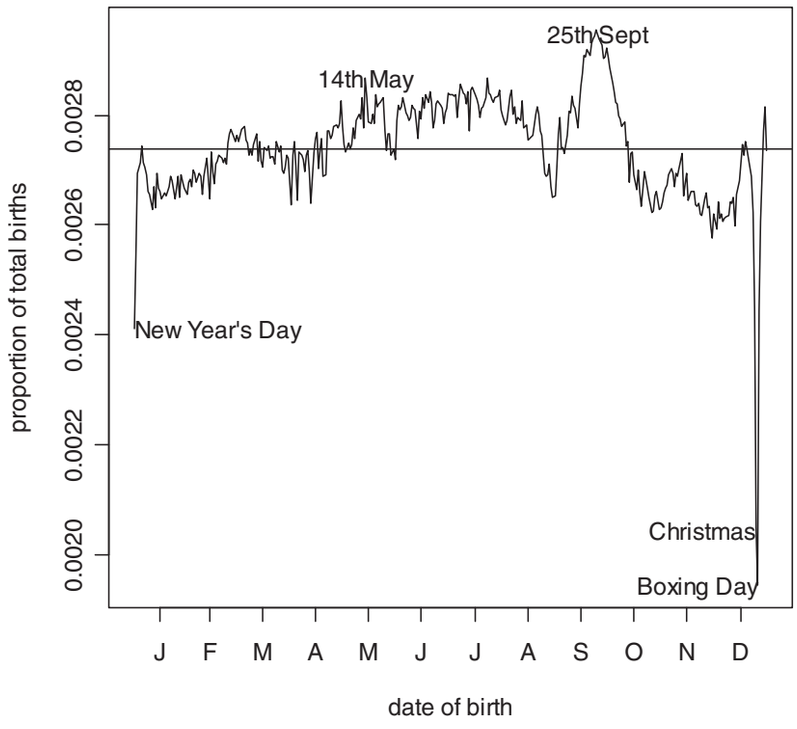


Figure 2. Daily proportions of births: England and Wales, 1979–2005; the horizontal line corresponds to a uniform distribution

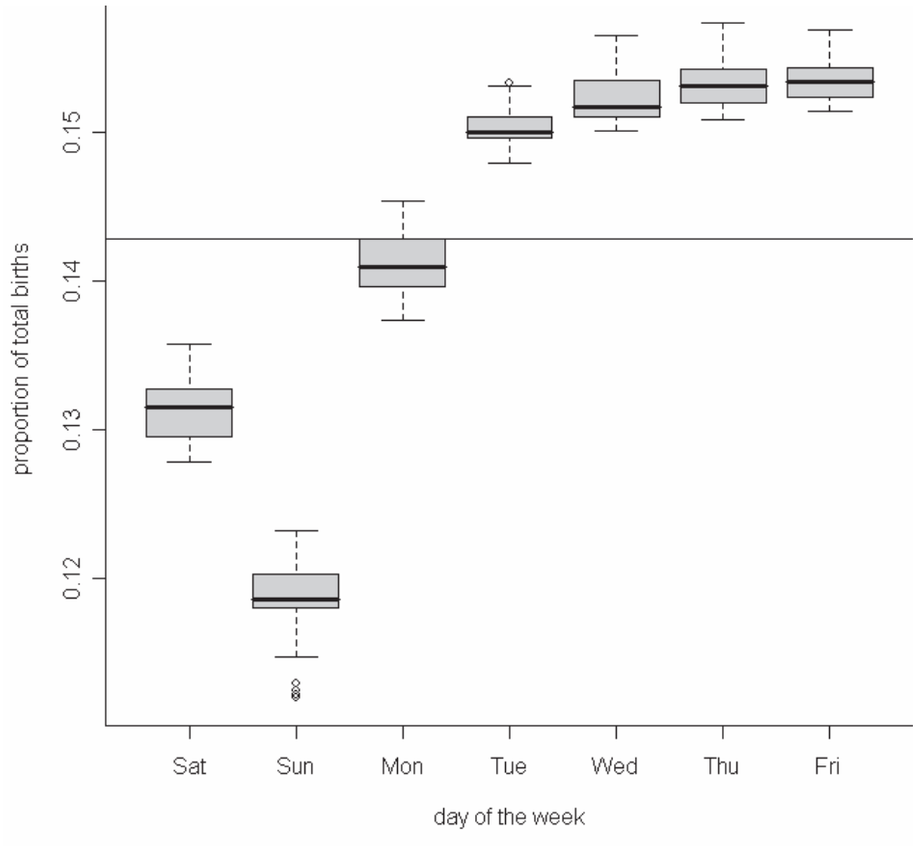


Figure 3. Boxplots of proportion of yearly births by day of the week: England and Wales, 1979 to 2005; the horizontal line corresponds to a uniform distribution (the Sunday outliers are for 1979–1983)

$D(365,22)$ is multiplied by 0.99917 and the product remains comfortably above 50%. The answer to the "birthday problem" is indeed 23.

But for a school teacher, assessing the chance of a shared birthday in a class of around 25 pupils *all born within a 12-month period*, this overall value of S is not relevant.

In a class of 88 people, it is more likely than not that three of them share a birthday

The pronounced weekly pattern of births will mean that the value of S for these pupils will be much larger. Figure 4 shows that, in any particular year, the increase is around tenfold, with S varying from 2.18×10^{-5} in 1991 to 3.98×10^{-5} in 1979. However, even this largest value leaves 23 as the critical size, since the factor multiplying $D(365,22)$ is then 0.991.

This modification may make a small difference for some of the group sizes shown in Table 1. If it is desired to have a 95% chance of a pair with a common birthday, Table 1 indicates that $r = 47$ people are needed, assuming that all 365 days are equally likely. But, if the members of the group are chosen at random from people born in 1979, then $r = 46$ will be enough for this target, and $r = 69$ gives a 99.9% chance!

More birthday problems

For many investigations of birthday coincidences, the simple model, which assumes that all N dates are equally likely, will be adequate. Variations on the original problem that have been addressed include the following:

- in a group of size r, how likely is it that at least two people have a birthday no more than d days apart?
- when a group of size k call out their birth dates in turn, how long to wait until some *other* group member hears their own birth date mentioned;
- in a collection of r boys and r girls, find the chance that a boy and a girl share a birthday.

Diaconis and Mosteller[3] used the equidistribution model to explain why many "coincidences" are far from unlikely. Suppose, for a group of people, there are h different types

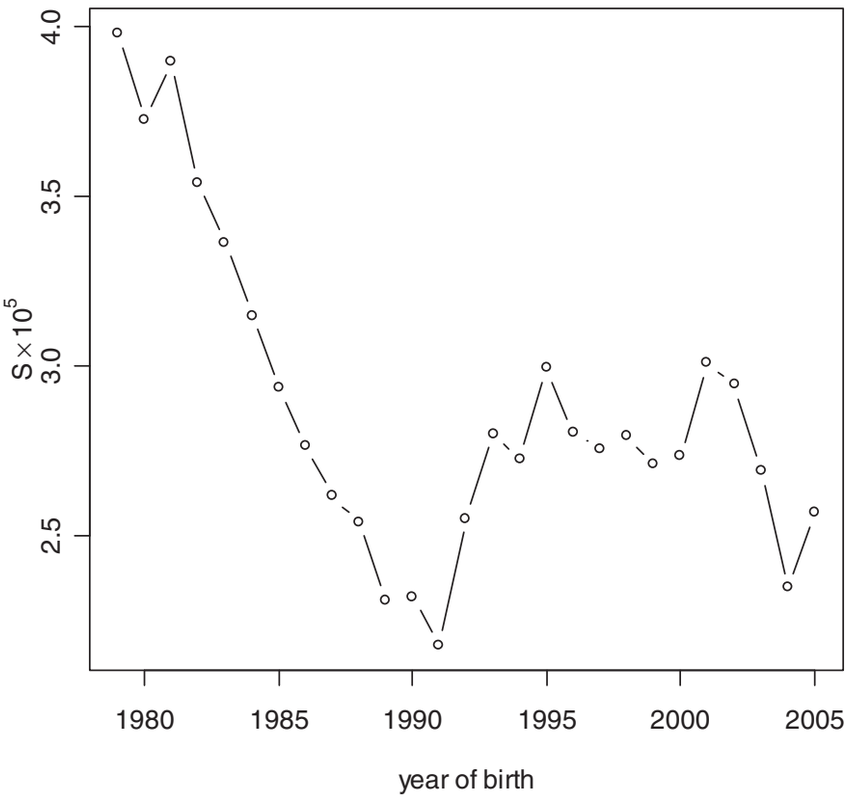


Figure 4. Empirical values of $S = \sum d_j^2 = \sum (p_j - 1/N)^2$ by year of birth: England and Wales, 1979 to 2005

of coincidence that one might meet—maybe birthdays, with $c_1 = 365$ possibilities, favourite two-digit number with $c_1 = 90$ choices, and so on. Then there will be an evens chance of a match in at least one of these categories provided that the group size is at least $1.2\sqrt{\{1/\sum_{i=1}^{b}(1/c_i)\}}$. Thus, for example, how many people are needed for an evens chance of a coincident birthday, either among the people themselves, or among their mothers or among their fathers? This formula indicates that as few as 13 suffice.

How large a group is required to make it more likely than not that at least *three* people share a birthday? For $1 \le i < j < k \le r$, write $X(i, j, k) = 1$ or 0 according to whether or not individuals $\{i, j, k\}$ share a birthday, and let $T(r) = \sum X(i,j,k)$. Plainly, its mean value is $\binom{r}{3}/N^2$, so, if the distribution of $T(r)$ is well approximated by a Poisson distribution, the critical value of r comes from $r(r-1)(r-2) = 6N^2\ln(2)$. This approach to the original problem for *two* to share a birthday leads to r being the solution of $r(r-1) = 2N\ln(2)$, which correctly gives $r = 23$ when $N = 365$, and is an excellent approximation whatever the value of N. In the present question about triples, we are led to $r = 82$ when $N = 365$, but the exact calculation[4] shows that at least 88 people are needed. Our estimate is not seriously wrong, but it is disappointingly inaccurate. Why does this approach work for a double birthday, but not for a triple?

Even in the case of two people, the corresponding quantities $\{X(i, j)\}$ are not independent, but they *are* pairwise independent. Thus, here $E\{T(r)\} = r(r-1)/(2N)$, and, because all the covariances in the expansion of $\mathrm{Var}\{T(r)\}$ are zero, then $\mathrm{Var}\{T(r)\} = E\{T(r)\}$. $((N-1)/N)$: mean and variance are virtually identical for large N. But for triple birthdays, we do not even have pairwise independence—consider, for example, $X(1, 2, 3)$ and $X(1, 2, 4)$.

This time, enough of the covariances in the expression for $\mathrm{Var}\{T(r)\}$ are non-zero for their total to be non-negligible. $T(r)$ is the sum of $\binom{r}{3}$ terms, each with variance $1/N^2 - 1/N^4$; and each of these $\binom{r}{3}$ terms has covariance $1/N^3 - 1/N^4$ with $3(r-3)$ other terms. Thus

$$\mathrm{Var}\{T(r)\} = E\{T(r)\}\left(1 + \frac{3(r-3)}{N} - \frac{3(r-8)}{N^2}\right)$$

Hence, for $N = 365$ and $r = 82$, the variance of $T(r)$ is about 1.65 times its mean, so $T(r)$ is *overdispersed* compared with the Poisson distribution. A larger variance tends to indicate higher probabilities in the tails of distributions—and the only value in the left tail of $T(r)$ is *zero*, as our choice of r makes its mean around ln(2). We should expect a Poisson approximation to underestimate the chance that $T(r) = 0$; hence the smallest value of r to ensure that this chance is below 0.5 will be higher than the Poisson approximation predicts—just as we have seen.

The same phenomenon will also occur for quadruple or higher order coincidences; the correct group size will be systematically higher than indicated by the Poisson approach.

Conclusions

In the language of balls and cells, it is often surprising how few "balls" are needed to get a reasonable chance of some form of "coincidence". The Poisson approximation to the distribution of the number of double birthdays in a group of size r can be expected to work well, but needs an adjustment for higher order coincidences.

But, whatever the problem, if there is an objection that the proposed model, taking all cells as equally likely, does not correspond to the real world, there is a devastating reply: invoking this extra touch of reality will mean that *even fewer* balls –or boys and girls in a class – will produce the desired result.

References

1. Munford, A. G. (1977) A note on the uniformity assumption in the birthday problem. *The American Statistician*, **31**, 119.

2. Nunnikhoven, T. S. (1992) A birthday problem solution with non-uniform birth frequencies. *The American Statistician*, **46**, 270–274.

3. Diaconis, P. and Mosteller, F. (1989) Methods for studying coincidences. *Journal of the American Statistical Association*, **84**, 853–861.

4. McKinney, E. (1966) Generalised birthday problem. *American Mathematical Monthly*, **73**, 385–387.

5. von Mises, R. (1964) *Selected Papers of Richard von Mises*, vol. 2, pp. 313–334. Providence. American Mathematical Society.

Mario Cortina Borja (m.cortina@ich.ucl.ac.uk) is Senior Lecturer in Statistics at the Institute of Child Health, University College London. John Haigh (j.haigh@sussex.ac.uk) is Reader in Mathematics at Sussex University. John was born on December 31st, 1941, the same day as Sir Alex Ferguson. Mario shares his birthday, April 19th, with tennis star Maria Sharapova, and with Richard von Mises (1883–1953), the pioneering statistician who first stated and solved the birthday problem.