The Birthday Problem Revisited

Untrue statements abound in my teaching. Some are unconscious, ranging from minor lapses of concentration, to errors that are simply arithmetical. Some represent a glitch in my thinking that has never been exposed before. Sometimes the untruth may be conscious: a simplification of the truth, because the more exact truth is maybe too difficult for this class at this time. I believe Picasso once described art as "a lie that makes us realise the truth." Perhaps the same could be said of teaching, at least in part.

Here is a lesson that proceeds smoothly apart from one tricky moment, where an assumption, which turns out to be reasonably justified, makes everything much easier. It is an assumption that many of my students would make without a thought. Could I use this as a pedagogic tool? Students are sometimes too ready to accept what appears on the whiteboard as gospel. Could they be warned that the following lesson will contain an error, deliberately inserted, and that one of their jobs is to identify it?

One lesson that has a safe place in the teaching calendar for me is the Birthday Problem, tackled fairly early on in the study of probability. This is the simple and hugely famous question, how many people do you need in a room for the probability to be at least 0.5 that some pair of them will have the same birthday? I usually ask students to consider this for a while, then offer them the following ranges, and ask for votes as to where the answer lies:

2-50, 51-100, 101-150, 151-200, 201-250, 251-300, 301-350

Then, by asking questions of increasing difficulty about dice, we home in on the answer from a theoretical point of view.

What is P(two 6s) when two dice are rolled?
What is P(no 6s) when two dice are rolled?
What is P(at least one 6) when two dice are rolled?
What is P(at least one 6) when n dice are rolled?

Out of this comes the idea that for n people,
P(at least one pair of people sharing a birthday)
= 1 – P(everybody has a different birthday)

$$= 1 - \frac{364 \times 363 \times 362 \times ... \times (365 - n + 1)}{365 \times 365 \times ... \times 365}$$

With the help perhaps of a graphics calculator program, it is not far to the answer: n=23. We return to the earlier voting, inevitably to find a mean estimate of many times that. The result never fails to startle. I decided to try a variant on this celebrated

problem with students entering the second year of their Statistics at A-Level (age 17-18). I started with the following situation.

Suppose that there are 365 people in the room. Clearly the chances that there are two people with the same birthday will be fantasically close to 1. So let's change the question:

What is the largest value of n such that the probability that there is a group of n people in the room with the same birthday is greater than or equal to 0.5?

This became s nice example of the Poisson approximation to the binomial. We supposed that the number of people with a birthday on a particular day was the random variable X. Now X ~ B(365, 1/365), which is close to the Poisson distribution with mean 1. What assumptions were we making here? Certainly that n = 365 is large and p = 1/365 is small, which seemed more than reasonable. We were also saying (as of course we say for the basic Birthday Problem) that every day is equally likely as a birthday, which is not quite true. In addition we were (again) agreeing to ignore February 29th as a birthday. Would twins or triplets be especially likely at this gathering for some obscure reason, and what effect would this have? After useful discussion on this, we turned to our cumulative Poisson tables, and found $P(x \leq 2) = 0.9197$ when $\lambda = 1$.

Now, (and here is the deliberate error),

This gives $P(x \leq 2)$ for all 365 days $= (0.9197)^{365} = 5.38 \times 10^{-14}$

So P(at least one group of three or more with the same birthday) $= 1 - (5.38 \times 10^{-14})$

The failing here is to treat the 365 Poisson distributions, all mean 1, as though they are independent, where they are not: their sum must come to 365 for a start. How big an error does this introduce? Can this be justified as an approximation? The best that I could eventually do to check this out was to run a computer simulation of the problem, which picks 365 birthdays at random and then finds the biggest hit, so to speak. If the program repeats this say 100 times, then the median of the biggest hit value becomes a good measure of n. It is possible to confirm that the error in taking the approximation as we did here was minute: the two methods agreed almost exactly.

Back in the lesson, tacitly accepting the method, we continued to look at our cumulative tables:

$$P(X \leq 4) = 0.9963,\ 0.9963^{365} = 0.258$$
$$P(X \leq 5) = 0.9994,\ 0.9994^{365} = 0.803$$

The fact that $^{365}\sqrt{0.5}$ is 0.99810 is helpful.

So P(there is at least one group of five or more with the same birthday) is 0.742 > 0.5, and P(there is one group of six or more with the same birthday) is 0.197 < 0.5. So we concluded that for 365 people in the room, n = 5: there is a probability of larger than 0.5 that there is a group of at least 5 people sharing a birthday, and 5 is the largest number for which this is true.

Can we generalise this? Calling the number of people in the room m, and considering first m = 183 ≈ 365/2, we arrive at the Poisson distribution with mean 0.5, and this gives us a value for n of 3. Choosing m = 3 × 365/2 gives the Poisson distribution with mean 1.5, and a value for n of 6, while m = 2 × 365 gives the Poisson distribution with mean 2, and a value for n of 7. Picking the values for m like this, a kind of graph starts to draw itself on our Poisson tables, as shown below.

	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5	5.5	6	6.5	7	7.5	8	8.5	9	9.5
1																			
2	0.9856																		
3	0.9982																		
4	0.9998	0.9963																	
5		0.9994	0.9955																
6		0.9999	0.9991	0.9955															
7			0.9998	0.9989	0.9958														
8				0.9998	0.9989	0.9962													
9					0.9997	0.9989	0.9967												
10						0.9997	0.999	0.9972											
11							0.9997	0.9991	0.9976										
12								0.9997	0.9992	0.998	0.9955								
13									0.9997	0.9993	0.9983	0.9964							
14										0.9998	0.9994	0.9986	0.997						
15												0.9995	0.9988	0.9976					
16													0.9996	0.999	0.998	0.9963			
17														0.9996	0.9992	0.9984	0.997		
18															0.9997	0.9993	0.9984	0.9976	
19																	0.9995	0.9989	0.998
20																		0.9996	0.9991
21																			0.9996
22																			

The graph is not quite linear to start with, but is (perhaps) surprisingly linear as the numbers get larger. Fitting a straight line to λ = 25, 50, 75, ... 200 gives the equation 14.5 + 1.14λ = n. (r = 0.9998). If this can be extrapolated, this gives n = 327 for 100,000 people. The simulation gave 323.

So imagine next Cup Final day at Wembley. Of the 100,000 people there, the pigeon-hole principle tells us that there must be a group of at least 274 with the same birthday. We now know that there is a probability of about 0.5 that there will be a group of at least 323 with the same birthday. And of course, there is a probability of better than 0.5 that two of the people on the pitch, the players and the referee, will share a birthday.

So, not a lesson for the purist perhaps, but one with a good range of ad hoc methods and approximations, which we check at any stage against our computer simulation. We arrive at a formula that works well, although we haven't derived it theoretically. The situation was easy to understand, and the discussions on assumptions that ensued were valuable.

So did they spot the error? The students had not met the Birthday Problem in any form before, and they worked it through with interest. They accepted that there was an error lurking in the lesson, which maybe added an edge to their concentration. They went along with, indeed provided the argument at the dubious moment. However, when taken back later, one and then all the students were able to spot the assumption about independence. Lessons since have shown that I can now get away with less!

Now if m is the number of people in the world, there is a probability greater than a half that there is a group of size...

Jonny Griffiths, Paston College, 1999

www.jonny-griffiths.net