Calculating the theoretical probability for the birthday paradox

It's much easier to calculate three misses in a row than it is to separately calculate the probability of 1, 2 and 3 made shots.

We often use simulation to estimate probabilities because it is difficult to work out the theoretical probability, however we can also use simulation to support a theoretical calculation. For the birth month and birthday problems, we can apply what we've learnt when studying probability rules.

The logic of complementary probabilities

First, we can reframe the way we think of there being "at least one match". You might have noticed in running birth month trials that in some cases you had more than one match, e.g., if we have (3, 6, 7, 3, 3) or (2, 2, 7, 7, 12). Whether there are two pairs or three people with the same birth month, we count it as a single 'success' as long as there's at least one pair with the same birthday.

This might remind you of the problems involving a basketballer shooting free-throws. The probability of hitting *at least one* free-throw out of three is the **complementary probability** of hitting *no* free throws – and it's much easier to calculate three misses in a row than it is to separately calculate the probability of 1, 2 and 3 made shots.

As the second person joins, they just have to have a different birthday to the first person.

In the birthday and birth month problems, we can hence look at the probability of everyone in the sample being born in different months - i.e., what is the probability of **no matches**? The complement of this will be the probability that we're after.

Let's consider each person joining our sample as a probabilistic event.

As the first person joins, there is a 100% chance of there being no match (since there's only one person in the group so far).

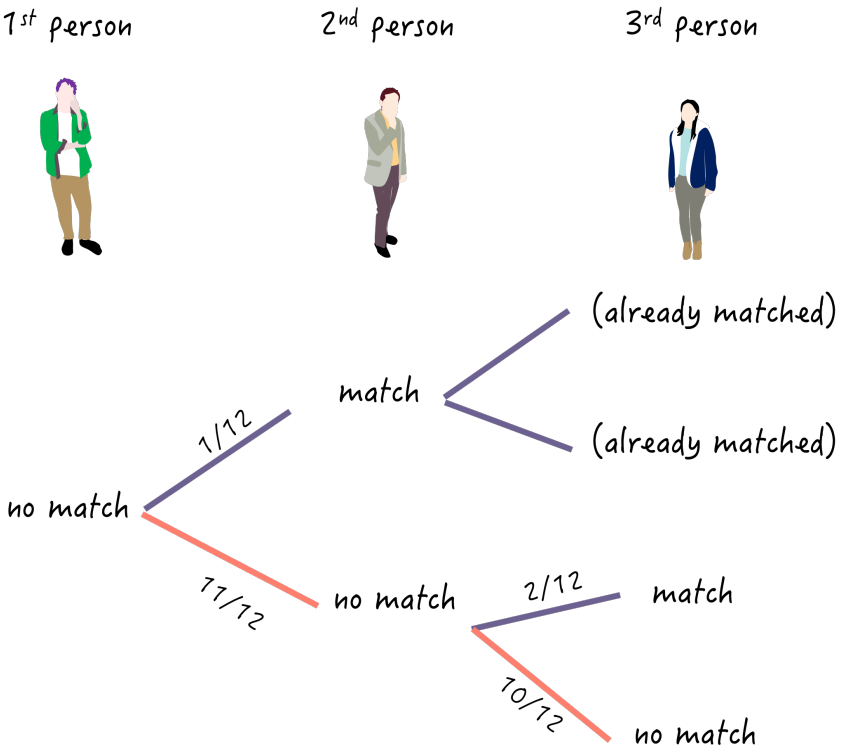
As the second person joins, they just have to have a different birthday to the first person. Whichever month the first person was born in, there'll be 11 months leftover. So the probability of there being no matches so far is

$$1 \times \frac{11}{12} = \frac{11}{12}.$$

When the third person joins, it needs to already be the case that the first two didn't match (which has probability 11/12) and then there will be 10 months leftover – as long as the third person is born in one of those 10 months, there still won't be a match. So now the probability is

$$1 \times \frac{11}{12} \times \frac{10}{12} = \frac{110}{144} = \frac{55}{72}.$$

We could follow the progress of this calculation using a tree diagram.



Then continuing on to the fourth and fifth person, we'll have 9 months and 8 months leftover respectively. The final calculation for the probability of **no matches** is hence:

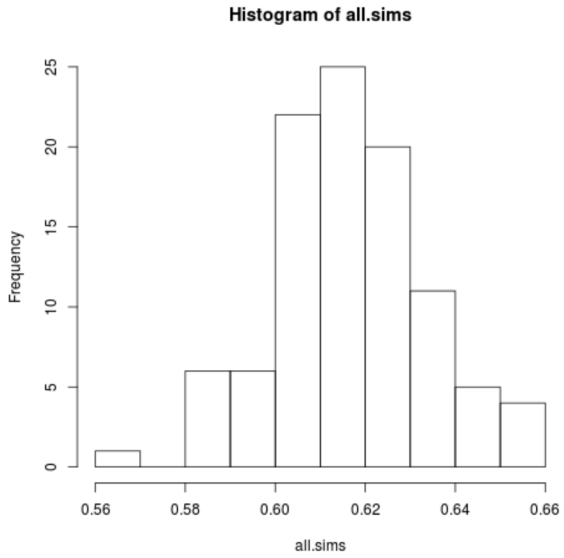
$$1 \times \frac{11}{12} \times \frac{10}{12} \times \frac{9}{12} \times \frac{8}{12} = \frac{7920}{20736} \approx 0.382\,.$$

Finally, to work out what the probability of there being at least one pair who are born in the same month, we can subtract the probability of no matches from 1. Hence we have 1 – 0.382 = **0.618** as our probability (correct to 3 decimal places).

We can apply the same logic to the birthday paradox.

How does this compare to our experimentation?

Using the R code from the unit site, the histogram you obtain (for 100 simulations of 1000 trials) might look something like this,



which does seem to centre around our theoretical calculation, even if there is some variation.

The birthday paradox

We can apply the same logic to the birthday paradox, however now we will have 365 days instead of 12 months, and 23 people instead of 5. The calculation will be:

$$1 \times \frac{364}{365} \times \frac{363}{365} \times \frac{362}{365} \times \ldots$$

and so on.

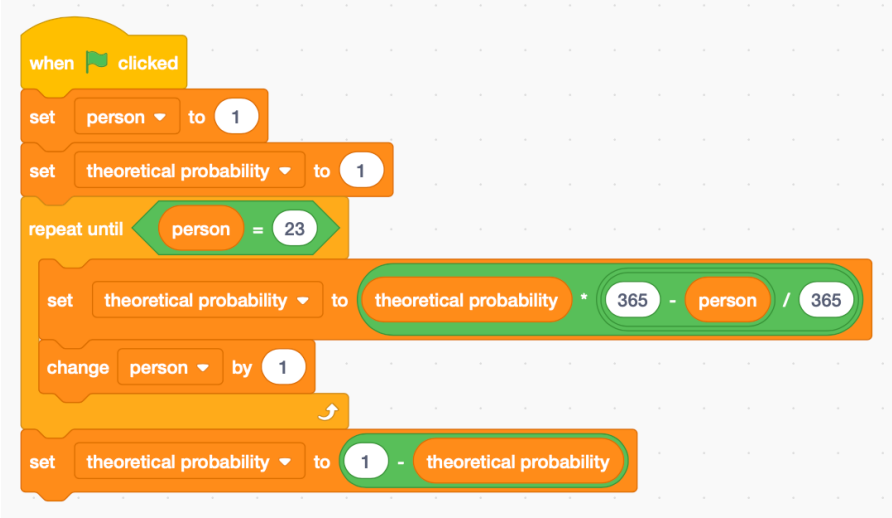
It might take you a little while to enter this into your calculator and reach the result, though.

As well as conducting our simulations, programming can be useful for theoretical calculations.

Programming to the rescue

As well as conducting our simulations, programming can also be useful for theoretical calculations such as the one above. The calculation is similar to what we'd have for a geometric progression, except that the ratio changes for each term.

In Scratch, the following code could be used.



(Scratch CC-BY-SA)

The probability starts as 1, then as each person enters the group, we multiply that probability by 365 minus the number of people already there, divided by 365. When **person = 22**, we'll be multiplying by (365 – 22) and then once person is increased to 23 the loop stops. At the end, we subtract this probability from 1.

To explore more, you might like to calculate the theoretical value for 10 people, 20 people and so on, then see how these compare to your simulated estimates.