tech space

Probability taught visually

The Birthday Paradox

Simulating using MS Excel

Jonaki Ghosh

One of the fundamental concepts in statistics is that of probability. It forms an integral part of most mathematics curricula at high school level. The topic of probability can be enlivened using many interesting problems. The related experiments are, however, time consuming and impractical to conduct in the classroom. Simulation can be an effective tool for modeling such experiments. It enables students to use random number generators to generate and explore data meaningfully and, as a result, grasp important probability concepts. This section discusses a well known problem known as the Birthday Problem or the Birthday Paradox which highlights an interesting paradox in probability and lends itself to investigation. Its exploration using a spreadsheet such as MS Excel can lead to an engaging classroom activity. It highlights the fact that spreadsheets can enable students to visualize, explore and discover important concepts without necessarily getting into the rigor of mathematical derivations.

Exploring the Birthday Paradox

The birthday problem, more popularly referred to as the birthday paradox, asks the following:

How many people do you need in a group to ensure that there are at least two people who share the same birthday (birth date and month)?

The immediate response from most students is 366 (which is true). However we find some surprises here. It can be shown that in a group size of 50 we can be *almost certain* to find a birthday match (often there is more than one match), and in group sizes of 24, the chance of finding a match is around half. The argument for this can lead to an interesting classroom discussion where basic concepts of probability play an important role.

It can be an interesting, although tedious, exercise to actually verify the claim empirically by randomly collecting birthdays, randomly dividing them into groups of 50 and checking if each group has a match. Another way of conducting the experiment is to ask each student to contribute 10 birthdays of persons known to her (relatives or friends), write them on slips of paper, fold them and put them in a box. After shaking the box, each student is asked to select a slip from the box and report the date which is then marked off on a calendar. The box is circulated till a date is repeated and number of dates that were marked before finding the match is noted. After performing this experiment several times the average number of dates required to find a match is calculated. Suppose 10 sets of 24 slips each are created from the contents of the same box, then students can verify that almost invariably 5 of the 10 sets will contain a match while the other 5 will not have a match. This helps to convince them that the probability of a match among 24 randomly selected persons is around half. While the exercise is exciting it can be very time consuming.

Simulation of the Birthday Problem on Excel

Simulating the problem on Excel, on the other hand, makes it far more convenient to do the experiment. 50 birthdays can be randomly generated using the **RAND( )** and **INT( )** functions.

Step 1: The first step is to randomly generate 50 integers between 1 and 12 (inclusive) in column A to indicate the months. This may be achieved in the following manner.

- Click on cell A2 and enter 1. Then enter **= A2 + 1** in cell A3 and drag cell A3 till A51. This will create a column of numbers 1 to 50 as shown in Figure 1.
- To generate 50 integers between 1 and 12 (inclusive) for indicating the months we enter **=INT (12 * RAND( ) + 1)** in cell B2. A double click on the corner of cell B2 will fill the cells B2 to B51 with 50 randomly generated integers between 1 and 12. These represent the months of 50 birthdays (as shown in Figure 1).

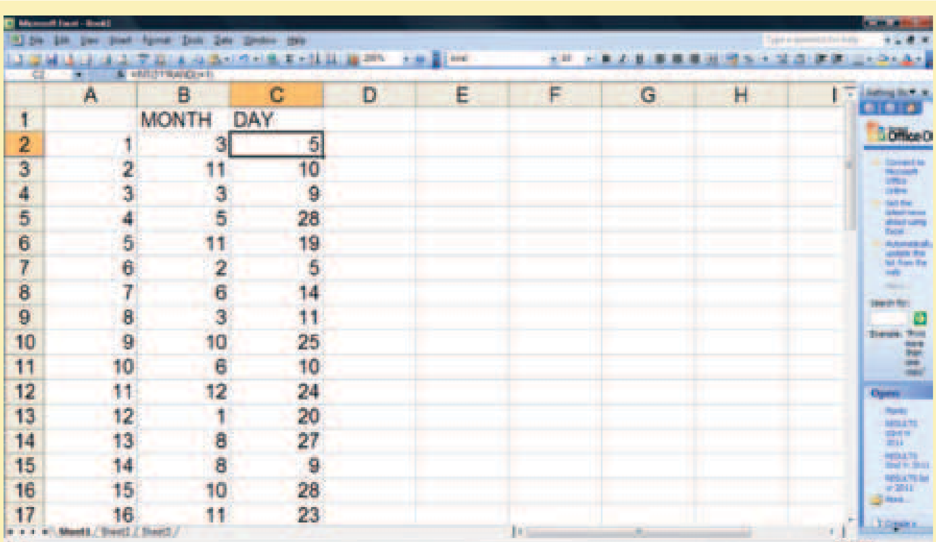


Figure 1: Simulation of birthdays in Excel

Step 2: The next step is to generate 50 random integers between 1 and 31 (inclusive) to indicate the day of the month. This is obtained as follows

1. Enter **=INT(31*RAND( )+1)** in cell C2 and double click on the corner of cell C2.
2. 50 randomly generated integers between 1 and 31 will appear in column C. These will represent days, corresponding to the months contained in column B.

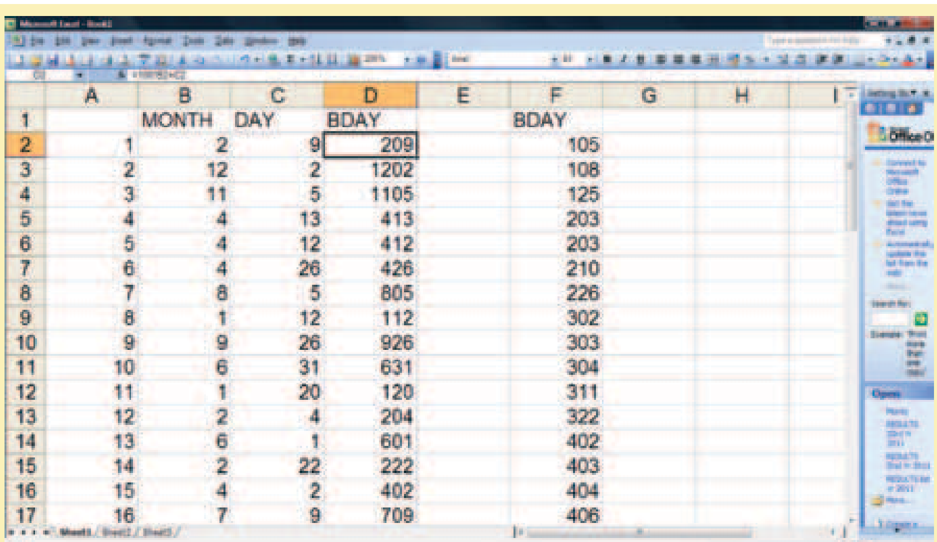
Step 3: The data in columns B and C represent 50 randomly generated birthdays. For example a 3 in cell B1 and 24 in cell C1 represent the date 24th March. We now need to browse through this list and search for a repeated date. This can be time consuming and inconvenient. In order to simplify this part of the process the dates may be converted to three or four digit numbers by entering the formula =100*B1+C1 in column D. Once this is done the first one or two digits of each number in column D will represent the month while the last two digits will represent the day. For example, the appearance of 225 in the list indicates 25th of February while 1019 indicates 19th of October. The list of numbers can then be arranged in an ascending order using the sorting feature of the spreadsheet. This will ensure that a repeated date will appear as two successive values and thus be easily identified.

To convert the dates in columns B and C to three or four digit numbers we enter **=100*B2+C2** in cell D3. Once again a double click on the corner of the cell D3 will reveal the 50 birthdays in cells D2 till D51 (see Figure 2).

Step 4: The list of dates appearing in column D needs to be sorted so that a birthday match can appear as two successive numbers and therefore be easily identified. To do this we select column D, go to **Edit**, select **Copy**, click on a column away from the data (for example choose a column from column F onwards), go to **Edit**, select **Paste Special** and click on **values** and then click on **OK**. This will ensure that all the numbers of column D will now be copied in the same sequence in the new column. Now click on the new column and select the sort (in ascending order) option from the toolbar. Once the dates are sorted a match can be easily identified as shown in Figure 2.

The experiment may be run about 10 times to confirm that in each simulation of 50 birthdays (representing the birthdays of 50 randomly selected people) there is at least one match. The pitfall of this simulation process is that impossible dates (such as 431, that is, 31st April etc) may appear in a particular list. In such a case the entire list can be ignored and the simulation may be repeated.

It might be useful, however, to follow the simulation exercise with an analysis of the problem using probability theory. Begin the discussion by finding the probability of a match in group sizes of three,



	A	B	C	D	E	F
1		MONTH	DAY	BDAY		BDAY
2	1	2	9	209		105
3	2	12	2	1202		108
4	3	11	5	1105		125
5	4	4	13	413		203
6	5	4	12	412		203
7	6	4	26	426		210
8	7	8	5	805		226
9	8	1	12	112		302
10	9	9	26	926		303
11	10	6	31	631		304
12	11	1	20	120		311
13	12	2	4	204		322
14	13	6	1	601		402
15	14	2	22	222		403
16	15	4	2	402		404
17	16	7	9	709		406

Figure 2: Column D represents 50 randomly generated birthdays. The same list is sorted in column F which represents a match (in this case 203, that is, 3rd February)

four and five. Once a pattern is evident, students can easily generalize it to find the formula for the probability of a match in a group size of n persons.

The probability that in a group of three persons, all three have distinct birthdays is

$$\frac{365}{365} \times \frac{364}{365} \times \frac{363}{365} = \frac{365 \times 364 \times 363}{365^3}.$$

Thus the probability that at least two of them share a birthday is

$$1 - \frac{365 \times 364 \times 363}{365^3}.$$

It needs to be emphasized here that in a group of three people there are three possible cases:

1. All three have distinct birthdays;
2. Two people have the same birthday while the third has a different birthday;
3. All three have the same birthday.

Since the three cases are mutually exclusive and exhaustive, the sum of their probabilities is 1. Thus the probability that at least two people have the same birthday includes cases (ii) and (iii) and can be obtained by subtracting the probability of (i) from 1.

The above expression can be extended to find the probability of at least one birthday match in a group of 4 persons, that is,

$$1 - \frac{365 \times 364 \times 363 \times 362}{365^4}.$$

Similarly, in a group of 5 persons the probability of a match is

$$1 - \frac{365 \times 364 \times 363 \times 362 \times 361}{365^5}$$

Extending this it can be shown that the probability of a match in a group of size n is

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

$$1 - \frac{365!}{(365 - n)! \times 365^n}$$

The value of the above expression approaches 1, as n approaches 50.

While generalizing the formula, students may need help in relating the last number of the product in the numerator to the group size, n. For example, the last number for $n = 3$ is $365 - 2 = 363$, for $n = 4$ it is $365 - 3 = 362$, for $n = 5$ it is $365 - 4 = 361$; for $n = k$, it is $365 - (k - 1)$. Once the generalized expression is obtained the knowledge of factorials may be used to write the expression in a concise manner.

Conclusion

The topic of probability has a plethora of interesting problems which can be made accessible to high school students through spreadsheets. The experiments related to these problems may be impractical to conduct manually but simulation can be an effective modeling tool for imitating such experiments. Microsoft Excel proves to be a very handy tool for conducting the explorations and investigations in the classroom. The Birthday problem discussed in this article can be conducted with students of grades 9 and 10 without getting into the mathematical derivations. However in grades 11 and 12 the spreadsheet verification of the problems can be followed by an analysis of the underlying concepts which are rooted in probability theory.

References

1. Microsoft Excel Training & Word 2007 Tutorial, Retrieved from http://www.free-training-tutorial.com/format-cells.html
2. Birthday Problem from Wikipedia, the Free Encyclopedia, Retrieved from http://en.wikipedia.org/wiki/Birthday_problem

JONAKI GHOSH is an Assistant Professor in the Dept. of Elementary Education, Lady Sri Ram College, University of Delhi where she teaches courses related to mathematics education. She obtained her Ph.D in Applied Mathematics from Jamia Milia Islamia University, New Delhi and Masters in Mathematics from Indian Institute of Technology, Kanpur. She has also taught mathematics at the Delhi Public School R K Puram for 13 years, where she was instrumental in setting up the Mathematics Laboratory & Technology Centre. She has started a Foundation through which she regularly conducts professional development programmes for mathematics teachers. Her primary area of research interest is in use of technology in mathematics instruction. She is a member of the Indo Swedish Working Group on Mathematics Education. She regularly participates in national and international conferences. She has published articles in proceedings as well as journals and has also authored books for school students. She may be contacted at jonakibghosh@gmail.com