Understanding the birthday problem

FREDERICK MOSTELLER, *Harvard University, Cambridge, Massachusetts.*

Comparing and contrasting the birthday problem and the birthmate problem

THE CLASSICAL BIRTHDAY PROBLEM: What is the least number of people required to assure that the probability that two or more of them have the same birthday exceeds $\frac{1}{2}$?

The usual simplifications are that February 29 is ignored as a possible birthday and that the other 365 days are regarded as equally likely birth dates. The problem and its answer, 23 people, are well known.[1]

Many are surprised at this answer, because 23 is a small fraction of 365. Instead, they feel that a number such as $183(\approx\frac{1}{2}(365))$ would be more reasonable. Their surprise stems in part from a misunderstanding of the statement of the problem. The problem they have in mind is *the birthmate problem:* What is the least number of people I should question to make the probability that I find one or more birthmates exceed $\frac{1}{2}$?

This problem is known, but not well known,[1] and its answer is 253. The answer exceeds 183 because the sampling of birth dates is done with replacement, rather than without. If we sampled from 365 people with *different* birth dates, 183 is the correct answer to the birthmate problem.

In comparing the birthday and birthmate problems, one observes that for r people in the birthday problem, there are $r(r-1)/2$ pairs or *opportunities* for like birthdays; whereas, if n people are questioned in the birthmate problem, there are only n opportunities for me to find one or more birthmates. It seems reasonable that, approximately, if one is to have about the same probability of success (finding matching birthdays) in the two problems, then $n\approx r(r-1)/2$, because then one has the same number of opportunities in both problems. For example, if the probability of success is to be at least $\frac{1}{2}$, then if $r=23$, $n=23(22)/2=253$, which happens to be exactly the correct answer. On the other hand, the probabilities of success in the two problems are not identical. For the birthday problem with $r=23$, $P(\text{success})\approx 0.5073$, for the birthmate problem $n=253$, $P(\text{success})\approx 0.5005$.

Let us generalize the two problems. Let N be the number of equally likely days (or categories), r the number of individuals in the birthday problem, and n the number examined in the birthmate problem. Then for the birthday problem it is easiest to compute first the probability of no like birthdays, and to obtain the probability of at least one like pair by taking the complement.

There are N ways for the first person to have a birthday, $N-1$ for the next so that he does not match the first, $N-2$ for the third so that he matches neither of the first two, and so on down to $N-r+1$ for the rth person. Then applying the multiplication principle, the number of ways for no matching birthdays is

$$N(N-1)\cdots(N-r+1), \quad r \text{ factors.} \qquad (1)$$

[1] W. Feller, *Probability Theory and Its Applications* (New York: John Wiley & Sons, Inc., 1950 [first edition]), I, pp. 31–32, 112.

To get the probability of no matching birthdays we also need the number of ways r people can have birthdays without restriction. There are N ways for each person. Then the multiplication principle says that the total number of different ways the birthdays can be assigned to r people is

$$N^r. \qquad (2)$$

The number in expression (1) divided by that in expression (2) is the probability of no like birthdays, because we assume that all birthdays and therefore all ways of assigning birthdays are equally likely. The complement of this ratio is the probability, P_R, of at least one pair of like birthdays in the classical birthday problem.

Thus

$$P_R = P(\text{at least 1 matching pair}) = 1 - N(N-1) \cdots (N-r+1)/N^r. \qquad (3)$$

For the birthmate problem the probability that a randomly chosen person does not have my birthday is $(N-1)/N = 1-1/N$. If n people are independently drawn, then by the multiplication principle for independent events, the probability that none has my birthday is $(1-1/N)^n$. The probability, P_S, that at least one has my birthday is

$$P_S = 1 - \left(1-\frac{1}{N}\right)^n. \qquad (4)$$

We wish to equate P_S to P_R approximately, and thus approximate n as a function of r. If $P_R = P_S$, then

$$\frac{N(N-1)(N-2) \cdots (N-r+1)}{N^r} = \left(1-\frac{1}{N}\right)^n. \qquad (5)$$

Dividing each factor on the left by an N from N^r, we get

$$1\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right) \cdots \left(1-\frac{r-1}{N}\right) = \left(1-\frac{1}{N}\right)^n. \qquad (6)$$

If we multiply out the left-hand side of Equation (6) to terms of order $1/N$, and expand the right-hand side to two terms we get

$$1 - \left(\frac{1}{N}+\frac{2}{N}+ \cdots +\frac{r-1}{N}\right) + \cdots = 1 - \frac{n}{N} + \cdots. \qquad (7)$$

Eliminating the 1's shows that, approximately, we want

$$n = 1+2+ \cdots +(r-1) = (r-1)r/2. \qquad (8)$$

This result justifies in part our earlier statement about the need for equivalent opportunities.

The approach just given is satisfactory if n is small compared to N, but the following logarithmic approach leads to a more refined relation between n and r if $P_R = P_S$.

Recall that if $|x| < 1$, then

$$\log_e (1-x) = -\left(x+\frac{x^2}{2}+\frac{x^3}{3}+ \cdots\right). \qquad (9)$$

Taking the logarithm of both sides of Equation (6) gives us

$$\log\left(1-\frac{1}{N}\right) + \log\left(1-\frac{2}{N}\right) + \cdots + \log\left(1-\frac{r-1}{N}\right) = n \log\left(1-\frac{1}{N}\right). \qquad (10)$$

We use Equation (9) to expand each logarithm in Equation (10) to get (after multiplying both sides of the equation by -1) Equation (11). Each line on the left side of Equation (11) corresponds to one logarithmic expansion in the left of Equation (10).

$$\left.\begin{array}{l} \dfrac{1}{N}+\dfrac{1}{2N^2}+\dfrac{1}{3N^3}+\cdots \\ +\dfrac{2}{N}+\dfrac{2^2}{2N^2}+\dfrac{2^3}{3N^3}+\cdots \\ +\dfrac{3}{N}+\dfrac{3^2}{2N^2}+\dfrac{3^3}{3N^3}+\cdots \\ +\cdots \\ +\dfrac{r-1}{N}+\dfrac{(r-1)^2}{2N^2}+\dfrac{(r-1)^3}{3N^3}+\cdots \end{array}\right\} = n\left(\frac{1}{N}+\frac{1}{2N^2}+\frac{1}{3N^3}+\cdots\right). \qquad (11)$$

Now we sum the left side of Equation (11) column by column, using the usual formulas for the sum of the integers, the sum of the squares of the integers and so on to get

$$\frac{r(r-1)}{2N}+\frac{r(r-1)(2r-1)}{6(2N^2)}+\frac{r^2(r-1)^2}{4(3N^3)}+\cdots$$
$$= n\left(\frac{1}{N}+\frac{1}{2N^2}+\frac{1}{3N^3}+\cdots\right). \qquad (12)$$

To get a refined estimate of n in terms of r, we divide both sides of Equation (12) by

$$\frac{1}{N}+\frac{1}{2N^2}+\frac{1}{3N^3}+\cdots,$$

and by ordinary but arduous long division we find

$$\begin{array}{r|l} & \dfrac{r(r-1)}{2}+\dfrac{r(r-1)(r-2)}{6N}+\dfrac{r^2(r-1)(r-2)}{12N^2}+\cdots \\ \hline \dfrac{1}{N}+\dfrac{1}{2N^2}+\dfrac{1}{3N^3}+\cdots & \dfrac{r(r-1)}{2N}+\dfrac{r(r-1)(2r-1)}{12N^2}+\dfrac{r^2(r-1)^2}{12N^3}+\cdots \\ & \dfrac{r(r-1)}{2N}+\dfrac{r(r-1)}{4N^2}+\dfrac{r(r-1)}{6N^3}+\cdots \\ \hline & \dfrac{r(r-1)(r-2)}{6N^2}+\dfrac{(r+1)r(r-1)(r-2)}{12N^3}+\cdots \\ & \dfrac{r(r-1)(r-2)}{6N^2}+\dfrac{r(r-1)(r-2)}{12N^3}+\cdots \\ \hline & \dfrac{r^2(r-1)(r-2)}{12N^3}+\cdots \\ & \dfrac{r^2(r-1)(r-2)}{12N^3}+\cdots \\ \hline & +\cdots \end{array}$$

Thus factoring $r(r-1)/2$ out of the quotient gives

$$n \approx \frac{r(r-1)}{2}\left[1+\frac{r-2}{3N}+\frac{r(r-2)}{6N^2}\right]. \qquad (13)$$

It is amusing to note that the coefficient of $1/N$ is the number of triple birthdays, but I have not found an intuitive rationale for this.

Thus for $N=365$, $r=23$, to equate probabilities for the two problems, $n=258$. And for these numbers $P_S \approx 0.5073$ and $P_S \approx 0.5073$.

In the Table for the Birthday Problems, we give in column 1 a target probability. Column 2 gives for the birthmate problem the smallest value of n that provides a probability of success larger than the tar-

TABLE FOR THE BIRTHDAY PROBLEMS: $=N365$

(1) TARGET P	(2) n	(3) P_S	(4) r	(5) P_R	(6) n_1	(7) P_1	(8) n^*	(9) n_{13}
.01	4	.01091	4	.01636	6	.01633	6	6
.05	19	.05079	7	.05623	21	.05598	21	21
.10	39	.10147	10	.11695	45	.11614	45	45
.20	82	.20146	14	.22310	91	.22093	92	92
.30	131	.30190	17	.31501	136	.31141	138	138
.40	187	.40132	20	.41144	190	.40623	193	193
.50	253	.50048	23	.50730	253	.50048	258	258
.60	334	.60001	27	.62686	351	.61824	359	359
.70	439	.70013	30	.70632	435	.69682	447	447
.80	587	.80020	35	.81438	595	.80454	614	614
.90	840	.90019	41	.90315	820	.89456	851	851
.95	1092	.95001	47	.95477	1081	.94848	1128	1128
.99	1679	.99001	57	.99012	1596	.98746	1683	1682

get probability. Column 3 gives the corresponding probability of success, P_S. Column 4 gives the smallest value of r that produces a probability of success for the birthday problem greater than the target probability shown in column 1. The actual probability of success, P_R, is shown in column 5. The 6th and 7th columns show $n_1 = r(r-1)/2$, for the r given in column 4, together with the probability of success, P_1, in the birthmate problem, with n_1 people. Column 8 gives n^*, the value of n for the birthmate problem that makes the probability of success in the birthmate problem nearest to the probability of success, P_R, in the birthday problem. Column 9 gives n_{13}, the value of n obtained from the approximation of Equation (13). Over the range of target probabilities considered in the table, n, n_1, n^*, and n_{13} are rather close. Also the probabilities P_R and P_1 are close, though $P_R > P_1$.

1963 election

The Committee on Nominations and Elections is making plans for the 1963 election. Officers to be elected in 1963 are: president-elect, vice-president for college, vice-president for junior high school, and three directors.

The committee invites your suggestions for candidates. Please submit the names of able individuals, along with full information about their qualifications, to the chairman of the Committee on Nominations and Elections, Agnes Herbert, 4807 Lindsay Road, Baltimore 29, Maryland, not later than May 15, 1962.