The Birthday Problem Revisited

M. SAYRAFIEZADEH
Medgar Evers College of CUNY
Brooklyn, NY 11225

Introduction The birthday problem asks for the probability that at least two people in a group of k people will have the same birthday. The problem continues to attract interest in the classroom and its variations and generalizations provide context for further theoretical elaboration. References to some of this material are included at the end of this note. The present work was motivated by the need to provide an approximation formula for the solution of the birthday problem in a liberal arts course on the Nature of Mathematics. The main result enables students who have not yet studied calculus to approximate solutions to birthday-type problems.

If each person in a group of k people chooses a number at random from a given set of n numbers, the probability p that there will be at least one repetition is

$$p = 1 - \frac{P(n,k)}{n^k}, \quad \text{where } P(n,k) = \frac{n!}{(n-k)!} \quad \text{and} \quad k \le n. \tag{1}$$

For $n = 365$ the formula yields the solution for the birthday problem. The scientific calculators that most students have can calculate p for some n and k, but their range is quite limited. It is desirable to use approximation formulas that yield p to a reasonable degree of accuracy. One such formula is $p > 1 - e^{-k(k-1)/2n}$ [4, p. 33]. In this note we will derive another approximation that improves on this and is more elementary. The derivation is based on the relationship between the geometric and the arithmetic means of a set of positive numbers [7, p. 18]. We will also use Taylor's approximations to derive a formula for the upper bound of the error. That will be accessible to students with a knowledge of elementary calculus.

An elementary approximation formula for p We show that

$$p > 1 - \left(1 - \frac{k}{2n}\right)^{k-1}, \quad \text{for } k \le n. \tag{2}$$

The derivation involves first finding an upper bound for $q = P(n,k)/n^k$ from which the above lower bound for $p = 1 - q$ will follow. If the numerator and the denominator of q are written out and common terms are cancelled we obtain

$$q = \frac{P(n,k)}{n^k} = \frac{P(n-1,k-1)}{n^{k-1}} = \prod_{j=1}^{k-1}\left(1 - \frac{j}{n}\right). \tag{3}$$

Since the numbers in the product are not identical, the geometric mean is strictly less than the arithmetic mean. Taking the respective means yields the inequality

$$\left[\prod_{j=1}^{k-1}\left(1 - \frac{j}{n}\right)\right]^{\frac{1}{k-1}} < \frac{\sum_{j=1}^{k-1}\left(1 - \frac{j}{n}\right)}{k-1} = \frac{k - 1 - \sum_{j=1}^{k-1}\frac{j}{n}}{k-1} = \left(1 - \frac{k}{2n}\right). \tag{4}$$

The desired upper bound for q is given by

$$q = \prod_{j=1}^{k-1}\left(1 - \frac{j}{n}\right) < \left(1 - \frac{k}{2n}\right)^{k-1}. \tag{5}$$

For $k = 23$ in the birthday problem the exact value of p correct to four decimal places is $p = 0.5073$; the approximation formula yields $p > 0.5055$. As another example we may suppose that in a group of 55 people each person writes down a three-digit number independently from others. Here, for $k = 55$ and $n = 900$ the probability p of at least two people writing the same number is $p = 0.8144$, correct to four decimal places. The approximation formula gives a lower bound as $p > 0.8128$. The relatively small size of the error in these approximations is due to the near equality of the numbers $(1 - (j/n))$ involved in going from the geometric mean to the arithmetic mean.

To compare the usual approximation with this one, we define p_1, q_1, p_2, and q_2 by

$$p_1 = 1 - q_1 = 1 - \left(1 - \frac{k}{2n}\right)^{k-1}$$

and (6)

$$p_2 = 1 - q_2 = 1 - e^{-k(k-1)/2n}.$$

Then

$$\ln q_1 = (k-1)\ln\left(1 - \frac{k}{2n}\right) < -(k-1)\frac{k}{2n}.$$

The last inequality follows from $\ln(1-x) < -x$ for $|x| < 1$ (look at the Taylor expansion of $\ln(1-x)$ about the origin). And clearly,

$$q_1 < e^{-k(k-1)/2n} = q_2.$$

Thus $q < q_1 < q_2$ and

$$p > p_1 > p_2. \tag{7}$$

In the approximation of p, p_1 is an improvement over p_2.

The error in the approximation The error E in the approximation of p by p_1 satisfies the following inequalities

$$E = p - p_1 < p - p_2 < q_2(1 - e^{-\varepsilon}) < q_2\varepsilon, \qquad k \le n. \tag{8}$$

We will show that

$$\varepsilon < \frac{k^3}{6(n-k+1)^2} \tag{9}$$

by using a straightforward application of Taylor's linear approximation with an error term for $\ln(1-x)$ around the origin, when $0 < x < 1$. That is, $\ln(1-x) = -x - (x^2/2(1-\xi)^2)$ where $0 < \xi < x$. Applying this to $\ln q$ in (3) we have

$$\ln q = \ln \prod_{j=1}^{k-1}\left(1 - \frac{j}{n}\right) = \sum_{j=1}^{k-1}\ln\left(1 - \frac{j}{n}\right) = \sum_{j=1}^{k-1} - \frac{j}{n} - \frac{\left(\frac{j}{n}\right)^2}{2\left(1-\xi_j\right)^2},$$

where $0 < \xi_j < j/n \leq (k - 1/n)$. The first terms in the second sum add up to $-(k(k-1)/2n)$ and if we denote the rest by

$$\varepsilon = \sum_{j=1}^{k-1} \frac{\left(\frac{j}{n}\right)^2}{2(1-\xi_j)^2}, \tag{10}$$

we obtain $\ln q = -(k(k-1)/2n) - \varepsilon$ and

$$q = e^{-\frac{k(k-1)}{2n}} e^{-\varepsilon} = q_2 e^{-\varepsilon}. \tag{11}$$

The last inequality in (8) may be verified by looking at the error term in the linear Taylor's approximation of $f(x) = 1 - e^{-x}$ about the origin. This concludes the proof of (8).

To get the upper bound (9) for ε we use the condition $0 < \xi_j < (j/n) \leq (k - 1/n)$ to obtain

$$\frac{1}{(1-\xi_j)^2} < \frac{n^2}{(n-k+1)^2}$$

and apply it to ε in (10) to yield

$$\varepsilon < \sum_{j=1}^{k-1} \frac{j^2}{2(n-k+1)^2} = \frac{1}{2(n-k+1)^2}\left(\frac{k(k-1)(2k-1)}{6}\right).$$

Hence,

$$\varepsilon < \frac{k^3}{6(n-k+1)^2}.$$

For $n = 10^{10}$, $k = 10^5$ we have $p_1 = 0.3934670$, $p_2 = 0.3934663$, and $E < 1.01 \times 10^{-6}$. In this case the approximation formulas give p correct to five decimal places. In general, as we shall see below, the error in the approximations p_1 and p_2 tends to zero as n and k go to infinity.

The two terms on the right-hand side of the error inequality $E < q_2(1 - e^{-\varepsilon})$ in (8) are nonnegative and are bounded above by 1. If n is very large compared to k, then q_2 tends to zero, otherwise $(1 - e^{-\varepsilon})$ diminishes for large k. More precisely, if $n \leq k^{1.75}$, then

$$q_2 = e^{-(k(k-1)/2n)} \leq e^{-(k(k-1)/2k^{1.75})}.$$

For arbitrary $\varepsilon > 0$ we may choose $K_1 > 0$ so large that $e^{-(k(k-1)/2k^{1.75})} < \varepsilon$ for $k > K_1$. Then, $E < q_2(1 - e^{-\varepsilon}) < \varepsilon$.

On the other hand, for $n > k^{1.75}$, we have the inequalities

$$1 - e^{-\varepsilon} < 1 - e^{-(k^3/6(n-k+1)^2)} < 1 - e^{-(k^3/6(k^{1.75}-k+1)^2}.$$

The last difference tends to zero for large values of k. $K_2 > 0$ can now be chosen so large that for $k > K_2$ the right-hand term is less than ε. Letting $K = \max\{K_1, K_2\}$ and $N = K$ will assure that $E < \varepsilon$ for $k > K$ and $n > N$. Thus the error E goes to zero as n and k go to infinity.

It is interesting to note that replacing $E < q_2(1 - e^{-\varepsilon})$ by $E < q_2\varepsilon$ is not a huge loss even when ε is large. In that case q_2 will be "exponentially" small and will dominate the product.

A limiting behavior of p If k and n are related by $k = cn^{\alpha}$, where $c > 0$ and $\alpha > 0$, then p has an interesting behavior in the limit, as displayed below:

$$\lim_{n\to\infty} p = \begin{cases} 0 & \text{if } \alpha < \frac{1}{2} \\ 1 - e^{-\frac{c^2}{2}} & \text{if } \alpha = \frac{1}{2} \\ 1 & \text{if } \alpha > \frac{1}{2} \end{cases} . \tag{12}$$

Since the error in the approximation of p by p_2 is zero in the limit, it is sufficient to verify the above results for p_2. Let

$$k = cn^{\alpha} = cn^{\frac{1}{2}+\beta}$$

where the three cases are determined by taking β as positive, negative or zero. Making the substituting $k^2 = c^2 n n^{2\beta}$ in q_2 we have

$$\lim_{n\to\infty} q_2 = \lim_{n\to\infty} e^{-(k(k-1)/2n)} = \lim_{n\to\infty} e^{-(k^2/2n)} = \lim_{n\to\infty} e^{-(c^2 n^{2\beta}/2)}.$$

The conclusions in (12) readily follow from

$$\lim_{n\to\infty} p = \lim_{n\to\infty} (1 - q_2)$$

by examining three different cases of β.

The interesting case of

$$\lim_{n\to\infty} p = e^{-(c^2/2)} \quad \text{for } k = cn^{1/2}$$

enables us to estimate k for fixed n that will make the probability p a preassigned value. For example, to obtain $p > \frac{1}{2}$ we get an estimate for k as $k > \sqrt{2 \ln 2}\sqrt{n}$. In the birthday problem with $n = 365$ this formula gives $k > 22.49$. This is actually correct when k is taken as an integer. For $k = 23$, the exact probability is $p = 0.507$.

REFERENCES

1. M. Abramson and W. O. J. Moser, More birthday surprises, *Amer. Math. Monthly* 77 (1970), 856–858.
2. D. M. Bloom, A birthday problem, *Amer. Math. Monthly* 80 (1973), 1141–2.
3. M. Lawrence Clevenson and William Watkins, Majorization and the birthday inequality, this MAGAZINE, 64 (1991), 183–188.
4. William Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 1, 3rd edition, John Wiley & Sons, New York, 1970.
5. Gerald A. Heuer, Estimation in a certain probability problem, *Amer. Math. Monthly* 66 (1959), 704–706.
6. Robert L. Hocking and Neil C. Schwertman, An extension of the birthday problem to exactly k matches, *The College Math. J.*, 17 (1986), 315–321.
7. Nicholas D. Kazarinoff, *Geometric Inequalities*, Random House, New York, 1961.
8. E. H. McKinney, Generalized birthday problem, *Amer. Math. Monthly* 73 (1966), 385–387.