The Birthday Problem

D.Kessler and J.Schiff

1 The Problem

Assume there are D days in a year, and there is equal probability of being born on any particular day of the year. What is the probability that in a group of N people ($N \leq D$) there are (at least) two with the same birthday? (This problem is a mathematical abstraction of the real life question of finding the probability of two people sharing a birthday; for the real life problem we should also take into account effects such as leap years, that more people are conceived in the winter etc.)

2 Exact Solution

Call the probability that at least two share the same birthday P. Then $P = 1 - \tilde{P}$, where $\tilde{P}$ is the probability that all N people have different birthdays. $\tilde{P}$ is easy to calculate: There are D out of D choices allowed for the birthday of the first person, $D-1$ out of D for the second, $D-2$ out of D for the third, and so on, up to $D-N+1$ out of D for the Nth person. Thus,

$$\tilde{P} = \frac{D}{D} \cdot \frac{D-1}{D} \cdot \frac{D-2}{D} \cdots \frac{D-N+1}{D} = \frac{D!}{(D-N)!D^N} \ . \tag{1}$$

This formula, while exact, is inconvenient to work with for large values of D. It is difficult to see whether $\tilde{P}$ is large or small, for a given choice of N, D. If we are interested in large D (for real years $D = 365$ of course!), it would be useful to have some kind of expansion of $\tilde{P}$ in powers of $1/D$. Since $\tilde{P}$ is a product, it is actually easier to get such an expansion of $\ln \tilde{P}$. We write $\epsilon = 1/D$. Then

$$\tilde{P} = 1 \cdot (1-\epsilon) \cdot (1-2\epsilon) \cdots (1-(N-1)\epsilon) \ ,$$

and so

$$\ln \tilde{P} = \sum_{k=0}^{N-1} \ln(1-k\epsilon) \ .$$

Using the Taylor series for $\ln(1-x)$,

$$\ln(1-x) = -\sum_{j=1}^{\infty} \frac{x^j}{j} \ , \tag{2}$$

we have

$$\ln \tilde{P} = -\sum_{k=0}^{N-1} \sum_{j=1}^{\infty} \frac{(k\epsilon)^j}{j} = -\sum_{j=1}^{\infty} \frac{1}{j} \left(\sum_{k=0}^{N-1} k^j \right) \epsilon^j = -\sum_{j=1}^{\infty} \frac{S_j}{jD^j} \ , \tag{3}$$

where S_j is defined as the sum

$$S_j \equiv \sum_{k=0}^{N-1} k^j \ . \tag{4}$$

The first few S_j's are:

$$\begin{aligned}
S_1 &= \frac{N(N-1)}{2} \ , \\
S_2 &= \frac{N(N-\frac{1}{2})(N-1)}{3} \ , \\
S_3 &= \frac{N^2(N-1)^2}{4} \ , \\
S_4 &= \frac{N(N-1)(N-\frac{1}{2})(N^2-N-\frac{1}{3})}{5} \ , \\
S_5 &= \frac{N^2(N-1)^2(N^2-N-\frac{1}{2})}{6} \ .
\end{aligned}$$

Exponentiating (3), we have our second exact formula for $\tilde{P}$:

$$\tilde{P} = \exp\left(-\sum_{j=1}^{\infty} \frac{S_j}{jD^j} \right) \ . \tag{5}$$

Most of the rest of this article is about various approximations to $\tilde{P}$. The next section discusses approximations of (1) and (5) from a purely mathematical point of view. Section 4 then discusses the meaning of these approximations from a combinatorial point of view. This is important, as the same techniques can be applied in other problems, including many for which we do not have exact formulas to apply mathematical methods to. The final section deals with the link to the central limit theorem.

3 Mathematical Approximations

For large D, a very good approximation of $\tilde{P}$ is given by keeping only the first term in (5), giving

$$P \sim 1 - \exp\left(-\frac{N(N-1)}{2D} \right) \ . \tag{6}$$

In figure 1 we plot the exact answer for P and this approximation, both as functions of N, for $D = 365$. It evidently is a very reasonable approximation. In figure 2 we plot the error of the approximation as a function of N.

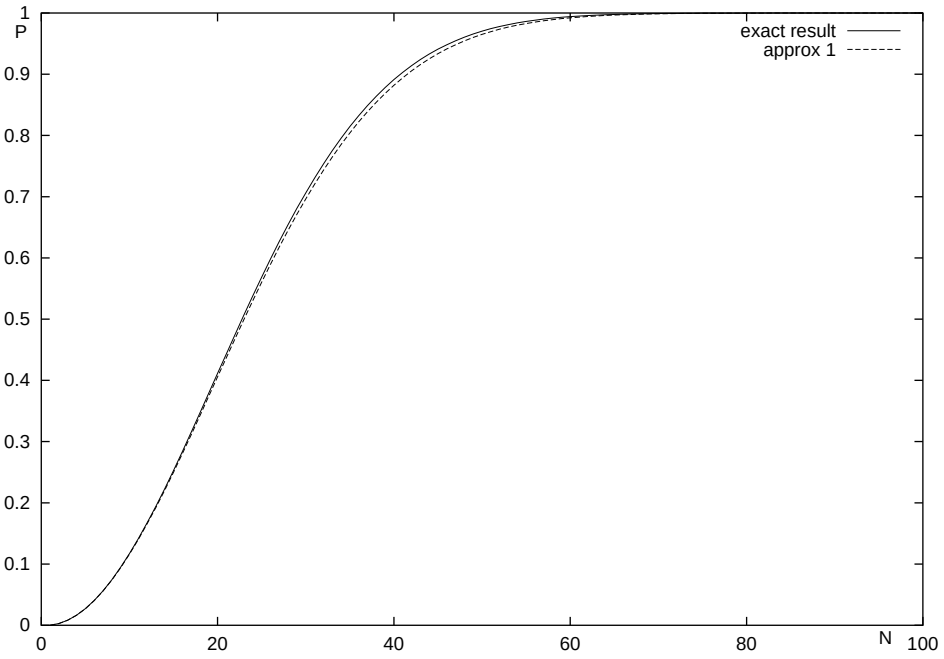

Figure 1: Exact answer for P and the approximation (6)

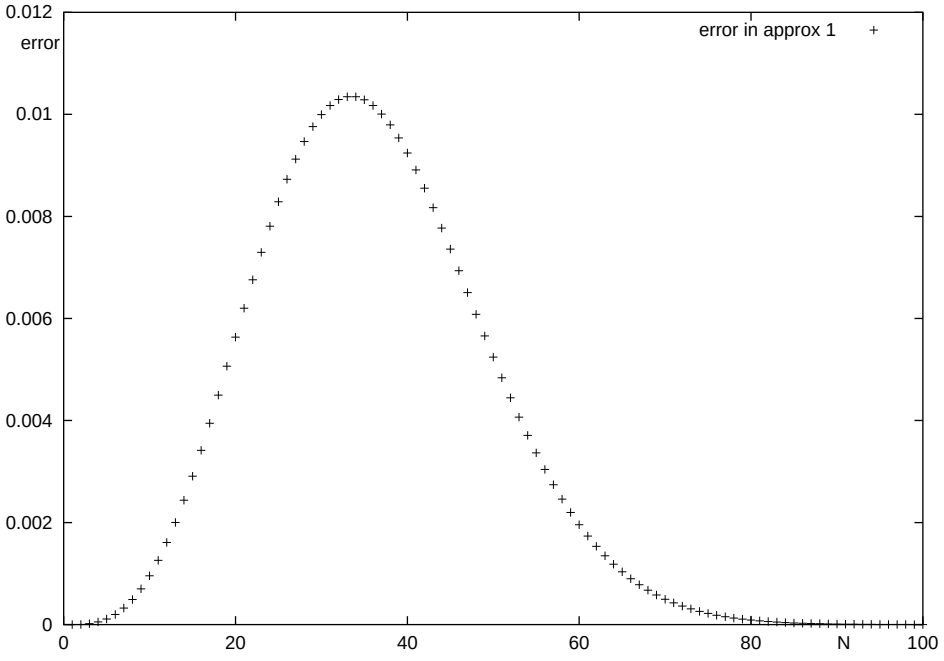

Figure 2: Error in approximation (6)

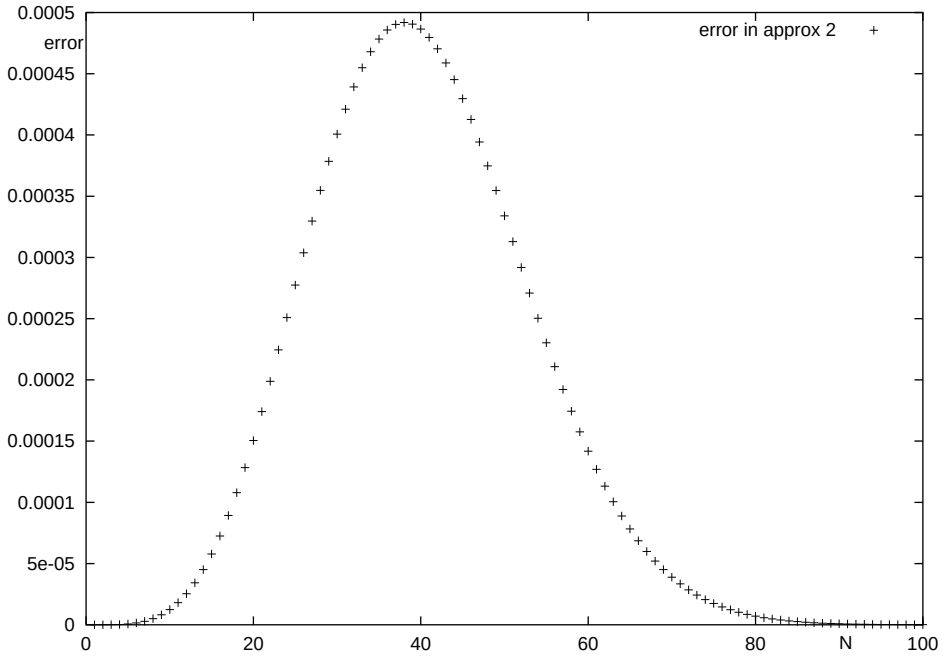

Figure 3: Error in approximation (7)

(6) implies that P reaches $\frac{1}{2}$ when $N(N-1)/2D \approx \ln 2$, or $N \sim \sqrt{2D \ln 2}$. To check the validity of this conclusion, we need to make sure the approximation we made is valid for such a choice of N. The second term in the series in (5) (the first term we neglected) is $S_2/2D^2$. When $N \sim \sqrt{2D \ln 2}$ this is of order $N^3/6D^2 = O(D^{-1/2})$. This is indeed still small, and neglecting it was justified. For $D = 365$ it is well-known that P first exceeds $\frac{1}{2}$ when $N = 23$. And indeed, $\sqrt{2 \times 365 \times \ln 2} \approx 22.5$.

If we wish, we can improve (6) by adding in further terms in the series in (5). The next approximation is

$$P \sim 1 - \exp\left(-\frac{N(N-1)}{2D} - \frac{N(N-1)(N-\frac{1}{2})}{6D^2}\right) \ . \tag{7}$$

In figure 3 we plot the error in this approximation as a function of N, for $D = 365$.

For large N the leading behavior of S_j is $N^{j+1}/(j+1)$. We can replace S_j by this in (5) and sum the resulting series to get

$$\tilde{P} \sim e^{-N}\left(1 - \frac{N}{D}\right)^{N-D} \ . \tag{8}$$

To get this you will need the result

$$\sum_{j=1}^{\infty} \frac{x^j}{j(j+1)} = 1 + \left(\frac{1}{x} - 1\right)\ln(1-x) \ ,$$

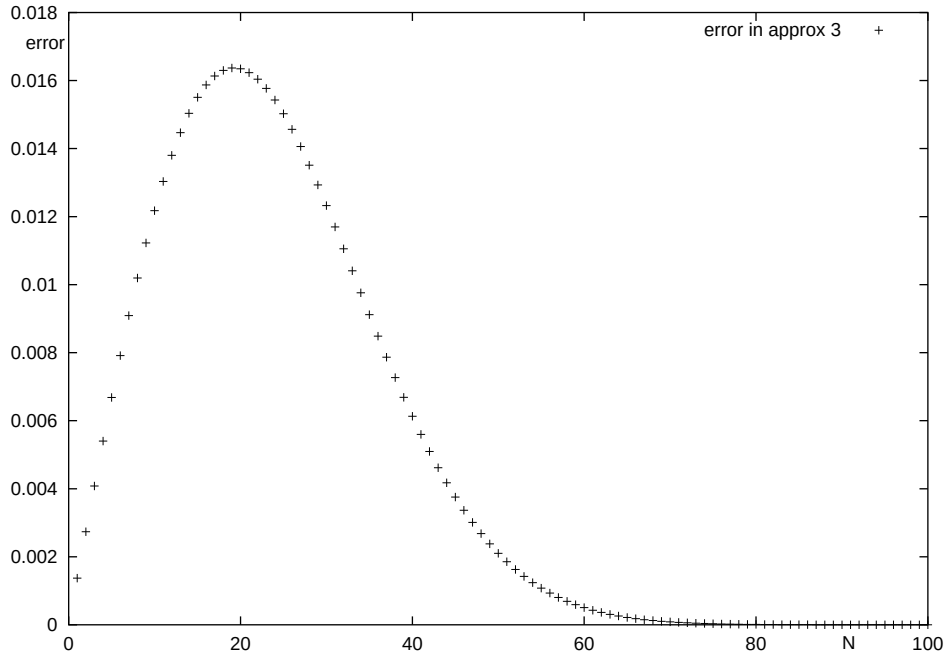

Figure 4: Error in approximation (8)

which can be obtained by integrating (2) with respect to x. The error in approximation (8) is shown in figure 4, again for $D = 365$. Approximation (8) can be painfully improved by taking more than just the leading behavior of S_j into account. But there is also another way of getting an improved version of (8). If we use Stirling's approximation $n! \approx \sqrt{2\pi n} n^n e^{-n}$ for the factorials in (1) we get

$$\tilde{P} \sim e^{-N} \left(1 - \frac{N}{D}\right)^{N-D-\frac{1}{2}} \quad . \tag{9}$$

In figure 5 we plot the error in this approximation. It is significantly more accurate than (6),(7) or (8).

4 Combinatorial Approximations

One "naive" approach to the original problem is as follows. In a group of N people there are $N(N-1)/2$ pairs, and the probability that a given pair shares the same birthday is $1/D$. Thus $P \approx N(N-1)/(2D)$. This is indeed the leading answer to order $1/D$, but it is not exact. Why?

The answer is that we cannot just treat all the pairs independently. Take the case of a

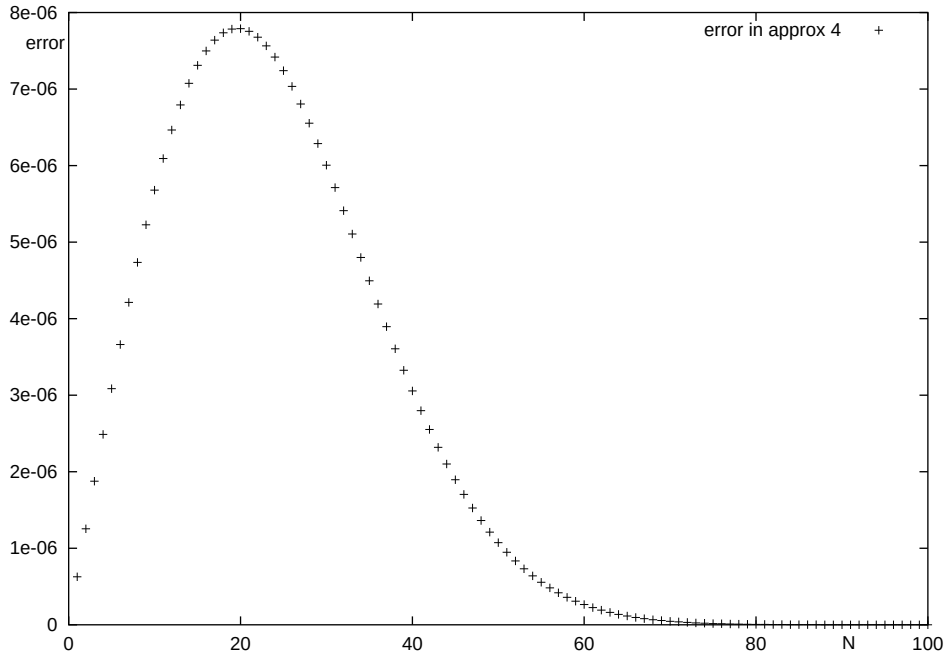

Figure 5: Error in approximation (9)

group of 3 people with 3 possible birthdays. There are 27 different possibilities:

111	112	113	121	122	123	131	132	133
211	212	213	221	222	223	231	232	233
311	312	313	321	322	323	331	332	333

(here "113" means the first and second person have birthday 1 and the third person has birthday 3). Of these in 9 cases persons 1 and 2 share a birthday, in 9 cases persons 2 and 3 share, and in 9 cases persons 1 and 3 share. If I naively add, I would predict that in all 27 cases some pair shares a birthday! This is wrong, because it overcounts the scenarios (111,222,333) in which all 3 share a birthday. Each of these scenarios gets counted 3 times, when it should only enter once. If I correct by this, I find that the number of cases in which some pair shares a birthday is $27 - 2 \times 3 = 21$ which is correct.

Unfortunately, in the general case it is not so simple to correct for the overcounting, as there are many different scenarios to take into account. But we can start the process. The first scenario is that there is a triple with the same birthday. The probability for this is $N(N-1)(N-2)/6D^2$, which we should subtract twice, as it is counted 3 times in the initial sum. Another possibility is 2 pairs, which happens with probability $3N(N-1)(N-2)(N-3)/24D^2$. Both these scenarios contribute terms of order $1/D^2$. It is easy to see that all other scenarios will contribute terms of higher order in $1/D$. Thus,

$$P \approx \frac{N(N-1)}{2D} - 2\frac{N(N-1)(N-2)}{6D^2} - \frac{3N(N-1)(N-2)(N-3)}{24D^2}$$

$$\approx \frac{N(N-1)}{2D} - \frac{N(N-1)(N-2)}{24D^2}[8+3(N-3)]$$
$$\approx \frac{N(N-1)}{2D} - \frac{N(N-1)(N-2)}{24D^2}(3N-1)$$

Expanding the exact solution (5) to order $1/D^2$ gives

$$\begin{aligned} P &\approx \frac{N(N-1)}{2D} - \frac{N^2(N-1)^2}{8D^2} + \frac{N(N-1)(N-\frac{1}{2})}{6D^2} \\ &\approx \frac{N(N-1)}{2D} - \frac{N(N-1)}{24D^2}\left[3N(N-1) - 4(N-\frac{1}{2})\right] \\ &\approx \frac{N(N-1)}{2D} - \frac{N(N-1)}{24D^2}(3N^2-7N+2) \\ &\approx \frac{N(N-1)}{2D} - \frac{N(N-1)}{24D^2}(3N-1)(N-2) \end{aligned}$$

which agrees with the above. Clearly this process can be extended, though very painfully. The leading order correction for large N comes from the 2 pair scenario. However, this is an overcorrection, because of the possibility of 3 pairs. How many are there? To leading order, $(5 \cdot 3 \cdot N^6)/6! = N^6/(2^3 3!) = (N^2/2)^3/3!$, which we need to add back in. Similarly with four pairs, etc. These terms exponentiate, giving $P \approx 1 - \exp(-N^2/2D)$, which is the leading result when N and D are big, with $N << D$.

The advantage of this approximate method is twofold: First, it gives the heart of the matter, the critical size of N immediately. Secondly, it can be extended to more complicated problems which do not admit a closed form solution. For example, how many people have to be in the group before 3 share a common birthday? The naive answer for the probability is $N(N-1)(N-2)/6D^2$ so that $N \sim (6D^2)^{1/3}$. How many people have to be in the group before there are two pairs? The naive answer for the probability is $3N(N-1)(N-2)(N-3)/24D^2$, so that $N \sim (8D^2)^{1/4}$, which is only a factor $2^{1/4}$ bigger than the critical number for one pair. Clearly, once we have enough people for one pair, the chance of two pairs is already nonneglible.

5 Connection to The Central Limit Theorem

The results for the birthday problem are the essence of the proof of the convergence of the binomial distribution $B(n,p)$ to a normal distribution for large n. Let us assume that np is an integer, so the mode of the binomial distribution occurs there. Recall that if $X \sim B(n,p)$

$$\mathbf{P}(X=r) = \binom{n}{r} p^r (1-p)^{n-r} \ .$$

So

$$\mathbf{P}(X = np) = \frac{n!}{(np)!(n(1-p))!} p^{np}(1-p)^{n(1-p)} \ .$$

Using Stirling's formula for all the factorials, this gives (after a little work)

$$\mathbf{P}(X = np) \approx \sqrt{\frac{1}{2\pi np(1-p)}} \ .$$

We also have

$$\begin{aligned} \mathbf{P}(X = r) &= \mathbf{P}(X = np)\ p^{r-np}(1-p)^{np-r} \frac{(np)!}{r!} \frac{(n-np)!}{(n-r)!} \\ &= \mathbf{P}(X = np)\ \frac{(np)!}{(np)^{np-r} r!} \frac{(n-np)!}{(n-np)^{r-np}(n-r)!} \end{aligned}$$

The penultimate factor is precisely the birthday formula for $\tilde{P}$ with $D = np$, $N = np - r$; the last factor is the birthday formula for $\tilde{P}$ with $D = n - np$, $N = r - np$ (you should have no cause to worry that in one factor N will be negative, all results remain true). Using approximation (6) for the birthday formula we have

$$\begin{aligned} \mathbf{P}(X = r) &\approx \mathbf{P}(X = np)\ \exp\left(-\frac{(np-r)(np-r-1)}{2np}\right) \exp\left(-\frac{(r-np)(r-np-1)}{2(n-np)}\right) \\ &= \mathbf{P}(X = np)\ \exp\left(-\frac{(np-r)^2 + (np-r)(2p-1)}{2np(1-p)}\right) \\ &= \mathbf{P}(X = np)\ \exp\left(\frac{(p-\frac{1}{2})^2}{2np(1-p)}\right) \exp\left(-\frac{(np+p-\frac{1}{2}-r)^2}{2np(1-p)}\right) \end{aligned}$$

The last term contains all the r dependence. If $p = \frac{1}{2}$ we have a Gaussian centered at $r = np$. If $p \neq \frac{1}{2}$ the center of the Gaussian is moved by $|p - \frac{1}{2}|$. The width of the Gaussian is $\sqrt{np(1-p)}$, as expected.