Puzzle [June, 1997]
Coincident Birthdays

1. How many people must be present to give a 50% probability of having (at least) two coincident birthdays in one year?

2. How many people must be present to give a 50% probability of having (at least) three birthdays in one year?

3. How many people must be present to give a 50% probability of having (at least) k coincident birthdays in one year, where k>3? How swiftly does this number grow with increasing k?

Mathcad 6.0 Solution by Patrice Le Conte (paraphrased by Steven Finch)

Solution for k=2

Assume that birthdays are independent and equiprobable. If $m := 365$ is the number of days in a year, there are a total of m possible outcomes for the first person, m^2 possible outcomes for the first two people, and thus m^p possible outcomes for the first p people.

Let H_1 be the number of all outcomes (out of m^p) where all people have different birthdays. There will be m possible birthdays for the first, $m - 1$ for the second, $m - 2$ for the third, and thus:

$$H_1 = \left[\prod_{i=0}^{p-1} m - i \right]$$

Therefore the probability that in a set of p people none have the same birthday is:

$$Q_1 = \frac{H_1}{m^p} \qquad Q_1(p) := \left[\prod_{i=0}^{p-1} 1 - \frac{i}{m} \right]$$

and the probability that at least two people have coincident birthdays is

$$P_2 = 1 - Q_1(p) \qquad P_2(p) := 1 - \left[\prod_{i=0}^{p-1} 1 - \frac{i}{m} \right]$$

$P_2(22) = 0.475695 \qquad P_2(23) = 0.507297$

The required number of people to have a 50% probability is:

$N_2 := 23$

Solution for k=3

We can use the same procedure to find the probability that the number of coincident birthdays is greater than two.

Let H_2 be the number of outcomes where the maximum number of coincident birthdays is exactly two. The probability of having a maximum of exactly two coincident birthdays in a set of p people is

$$Q_2 = \frac{H_2}{m^p}$$

and the probability of having at least three coincident birthdays is

$$P_3 = 1 - Q_2 - Q_1$$

Let C(m,n) be the number of combinations of m objects taken n at a time:

$$C(m,n) = \binom{m}{n} = \frac{m!}{n!\cdot(m-n)!} \qquad C(m,n) := \prod_{i=0}^{n-1} \frac{m-i}{n-i}$$

In order to compute H_2, we will separate the p people into two classes: one of $2\cdot i$ people whose birthdays are coincident, and one of $p - 2\cdot i$ people whose birthdays are not coincident.

First let us compute the number of outcomes where we have i coincident birthdays. We can choose $2\cdot i$ people out of p in $C(p, 2\cdot i)$ different ways. For each such selection of $2\cdot i$ people, there are

$$\frac{1}{i!}\cdot\binom{2\cdot i}{2}\cdot\binom{2\cdot i-2}{2}\cdot\binom{2\cdot i-4}{2}\cdot\ldots\cdot\binom{4}{2} = \frac{(2\cdot i)!}{i!\cdot 2^i} = \left[\prod_{j=1}^{i} 2\cdot j - 1\right]$$

different ways of arranging them into sets of two. Finally, each of the i pairs and the remaining $p - 2\cdot i$ people have distinct birthdays, and the number of ways this can happen is:

$$\left[\prod_{j=0}^{i-1} m - j\right]\cdot\left[\prod_{j=i}^{p-i-1} m - j\right] = \left[\prod_{j=0}^{p-i-1} m - j\right]$$

So the number of outcomes in which exactly $2 \cdot i$ people have coincident birthdays is:

$$H_{2_i} = C(p, 2 \cdot i) \cdot \left[\prod_{j=1}^{i} 2 \cdot i - 1 \right] \cdot \left[\prod_{j=0}^{p-i-1} m - j \right] = p! \cdot \frac{1}{2^i} \cdot C(m, i) \cdot C(m - i, p - 2 \cdot i)$$

Summing over i, we obtain:

$$H_2 = \sum_{i=1}^{\text{floor}\left(\frac{p}{2}\right)} H_{2_i}$$

$$Q_2 = \frac{p!}{m^p} \cdot \sum_{i=1}^{\text{floor}\left(\frac{p}{2}\right)} \frac{1}{2^i} \cdot C(m, i) \cdot C(m - i, p - 2 \cdot i)$$

which we rewrite in a way easier to compute:

$$Q_2(p) := \sum_{i=1}^{\text{floor}\left(\frac{p}{2}\right)} C(p, 2 \cdot i) \cdot \prod_{j=1}^{i} \frac{2 \cdot j - 1}{m} \cdot \left[\prod_{j=0}^{p-i-1} 1 - \frac{j}{m} \right]$$

The probability that at least three people have coincident birthdays is:

$$P_3(p) := 1 - Q_1(p) - Q_2(p)$$

$P_3(87) = 0.499455$ $\qquad$ $P_3(88) = 0.511065$

The required number of people to have a 50% probability is:

$N_3 := 88$

General Solution (all k)

We can use the same procedure to compute the general case of k people having the same birthday. First, the number of different ways of arranging $k \cdot i$ people into sets of k is:

$$\frac{1}{i!} \cdot \binom{k \cdot i}{k} \cdot \binom{k \cdot i - k}{k} \cdot \binom{k \cdot i - 2 \cdot k}{k} \cdot \ldots \cdot \binom{2 \cdot k}{k} = \frac{(k \cdot i)!}{i! \cdot (k!)^i}$$

Let $H(m,p,k)$ be the number of outcomes where the maximum number of coincident birthdays is exactly k. We first compute the number of outcomes where there are exactly i coincident birthdays of k people. This is done just as before, separating the p people into two classes: one of $k \cdot i$ people whose birthdays are coincident, and one of the remaining $p - k \cdot i$ people. There are

$$\left[\prod_{j=0}^{i-1} m - j\right]$$

ways each of the i sets can have distinct birthdays, and

$$\sum_{j=1}^{k-1} H(m-i, p-k\cdot i, j)$$

ways the remaining people can have birthdays (which needn't be distinct for k>2, hence the recursion). Therefore:

$$H(m,p,k)_i = C(p, k\cdot i)\cdot\frac{(k\cdot i)!}{i!\cdot(k!)^i}\cdot\left[\prod_{j=0}^{i-1} m - j\right]\cdot\sum_{j=1}^{k-1} H(m-i, p-k\cdot i, j)$$

Summing over i, we obtain:

$$H(m,p,k) = \sum_{i=1}^{\text{floor}\left(\frac{p}{k}\right)} H(m,p,k)_i = \sum_{i=1}^{\text{floor}\left(\frac{p}{k}\right)} C(p,k\cdot i)\cdot\frac{(k\cdot i)!}{i!\cdot(k!)^i}\cdot\left[\prod_{j=0}^{i-1} m - j\right]\cdot\sum_{j=1}^{k-1} H(m-i, p-k\cdot i, j)$$

Now we have $Q(m,p,k) = \frac{H(m,p,k)}{m^p}$ and thus $H(m-i, p-k\cdot i, j) = Q(m-i, p-k\cdot i, j)\cdot(m-i)^{p-k\cdot i}$ for all i.

Hence:

$$Q(m,p,k) = \sum_{i=1}^{\text{floor}\left(\frac{p}{k}\right)} \frac{C(p,k\cdot i)\cdot(k\cdot i)!}{m^{k\cdot i}}\cdot\frac{1}{i!\cdot(k!)^i}\cdot\left[\prod_{j=0}^{i-1} m - j\right]\cdot\sum_{j=1}^{k-1} Q(m-i, p-k\cdot i, j)\cdot\frac{(m-i)^{p-k\cdot i}}{m^{p-k\cdot i}}$$

which can be rewritten in a form better suited to computation as:

$$Q(m,p,k) = \sum_{i=1}^{\text{floor}\left(\frac{p}{k}\right)} \left(1 - \frac{i}{m}\right)^{p-k\cdot i}\cdot\prod_{j=1}^{k\cdot i}\frac{p-j+1}{m}\cdot\prod_{j=1}^{i}\frac{(m-j)+1}{j\cdot k!}\cdot\left(\sum_{j=1}^{k-1} Q(m-i, p-k\cdot i, j)\right)$$

Introduce, for convenience, a function:

$$q(m,p,k,i) := \left(1-\frac{i}{m}\right)^{p-k\cdot i}\cdot\prod_{j=1}^{k\cdot i}\frac{p-j+1}{m}\cdot\prod_{j=1}^{i}\frac{(m-j)+1}{j\cdot k!}$$

then we have the following recursive definition (note the initial conditions):

$$Q(m,p,k) := \begin{cases} 0 & \text{if } (p<k)+(m<1) \\ \text{otherwise} & \\ \quad \prod_{i=0}^{p-1}\left(1-\frac{i}{m}\right) & \text{if } k=1 \\ \quad \sum_{i=1}^{\text{floor}\left(\frac{p}{k}\right)} q(m,p,k,i)\cdot\text{if}\left(k\cdot i<p, \sum_{j=1}^{k-1} Q(m-i,p-k\cdot i,j), 1\right) & \text{otherwise} \end{cases}$$

So the probability that at least k people have coincident birthdays is:

$$P(m,p,k) := 1-\sum_{j=1}^{k-1} Q(m,p,j)$$

We confirm that $P(m,22,2) = 0.475695$, $P(m,23,2) = 0.507297$,
$P(m,87,3) = 0.499455$, $P(m,88,3) = 0.511065$

and compute that

$P(m,186,4) = 0.495826$ $P(m,187,4) = 0.502685$ $N_4 := 187$

As P is a recursive function, the time required for computation grows exponentially with k, so we merely record here the results for k=5 :

$P(m,312,5) = 0.496196$ $P(m,313,5) = 0.50107$ $N_5 := 313$

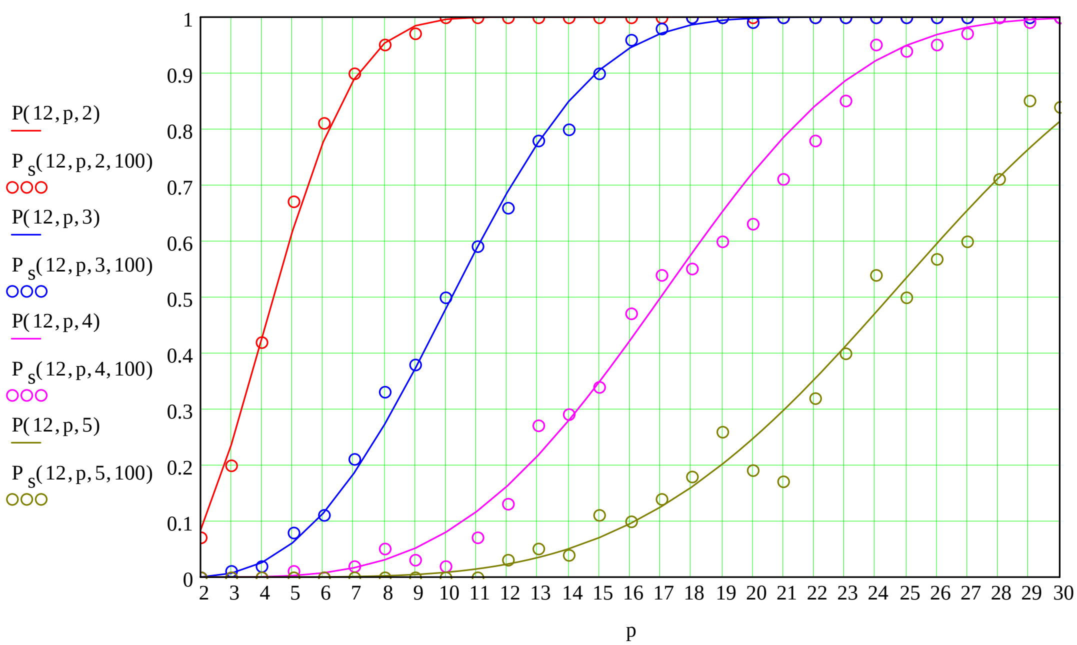
Let's try to verify these results through Monte Carlo simulation. The function $K_s(m,p)$ returns the maximum number of coincident birthdays in a set of p. The function $P_s(m,p,k,n)$ returns the probability of k coincident birthdates in a set of p, calculated by evaluating n times the function K_s and then counting how often its value exceeds k .

```
K_s(m,p) :=  | for i ∈ 0..m − 1
             |   a_i ← 0
             | for i ∈ 1..p
             |   | j ← floor(rnd(m))
             |   | a_j ← a_j + 1
             | max(a)
```

```
P_s(m,p,k,n) :=  | Σ ← 0
                 | for i ∈ 1..n
                 |   Σ ← Σ + 1  if  K_s(m,p) ≥ k
                 | Σ/n
```

It would take too much computation time to compare P and P_s for m=365, so we will use m=12, which can be interpreted as the number of coincident *months* of birth for a set of p people.

$p := 2..30$



This confirms the agreement between the calculated and simulated probabilities.

A remarkably accurate approximation, due to Bruce Levin [1], makes computations possible for larger k. See also [2, 3].

References

1. B. Levin, A representation for multinomial cumulative distribution functions, *Annals of Statistics* 9 (1981) 1123-1126.

2. P. Diaconis and F. Mosteller, Methods of studying coincidences, *J. Amer. Statist. Assoc.* 84 (1989) 853-861.

3. M. S. Klamkin and D. J. Newman, Extensions of the birthday surprise, *J. Combin. Theory* 3 (1967) 279-282.