The Birthday Problem

Tony Field

October 20, 2003

The Problem

Given m randomly selected people, let P_m be the probability that on each day of the year at least one of the m people has a birthday. We want to know the value of m for which $P_m > 0.5$. Assume birthdays are uniformly distributed over the year and that there are 365 days in the year (i.e. ignore leap years).

Facts

We know that if there are n days of interest (here 365 but we can choose any value we want), $P_m = 0$ if $m < n$, since if there are fewer people than days at least one day is "unallocated". We also know that

$$P_n = \frac{n!}{n^n}$$

since if there are exactly the same number of people as days then they must each have a different birthday.

Recurrence

Let $p_{k,m}$ be the probability that, assuming the number of days n is given, there are k unallocated days (i.e. with no birthdays) given m randomly-selected people. Then, $P_m = p_{0,m}$. We can establish the following recurrence:

$$\begin{aligned}
p_{n,0} &= 1 \\
p_{k,0} &= 0, \;\; k < n \\
p_{n,m} &= 0, \;\; m > 0 \\
p_{k,m} &= p_{k,m-1}\frac{n-k}{n} + p_{k+1,m-1}\frac{k+1}{n}, \;\; k < n, \, m > 0
\end{aligned}$$

Note for the latter: we get k unallocated 'gaps' with m people if either we get k gaps with $m-1$ people and the m^{th} person lands on one of the allocated days (probability $(n-k)/k$, or if we get $k+1$ gaps with $m-1$ people and the m^{th} lands on one of the gaps (probability $(k+1)/n$).

Observe that $p_{k,m} = 0$ if $k+m < n$. We expect this since each available day must either be allocated or not. From this we obtain:

$$\begin{aligned} p_{0,n} &= \frac{1}{n} \times p_{1,n-1} \\ &= \frac{1}{n} \times \frac{2}{n} \times p_{2,n-2} \\ &= \ldots \end{aligned}$$

which, together with the base case $p_{n,0} = 1$, is easily seen to produce the solution $p_{0,n} = n!/n^n$ as we would expect. A more rigorous proof by induction is straightforward.

Numerical Solution

We can treat p as a recursive function with appropriately defined base cases—a trivial Haskell program! The function will terminate for valid argument values but the computational complexity is exponential in n. We make the solution tractable by formulating it as a dynamic programming problem.

Imagine the parameter space $k, m, 0 \leq k \leq n, 0 \leq m \leq M$ as a matrix of size $n+1$ rows by $M+1$ columns, where M is the maximum number of people that we want to consider. The 0^{th} column and row are all zero with the exception of element $(n, 0)$ which is 1. These arise from the above base cases. Element $(k, m), 0 \leq k \leq n-1, 1 \leq m \leq M$ depends on the elements to the west $(k, m-1)$ and south-west $(k+1, m-1)$ so by iterating from element $(1, n-1)$ to element $(0, M)$, i.e. in a north-east direction, we obtain the solution in time proportional to $n \times M$.

Here is a selection of values from the table for $n = 365, M = 2290$ showing the point where we hit the break-even point. We conclude that the answer is 2287.

```
365,2280 solution is: 0.493657410104351
365,2281 solution is: 0.494618882351313
365,2282 solution is: 0.495579555664064
365,2283 solution is: 0.496539425608168
365,2284 solution is: 0.497498487775709
365,2285 solution is: 0.498456737785259
365,2286 solution is: 0.499414171281851
365,2287 solution is: 0.500370783936946
365,2288 solution is: 0.501326571448405
365,2289 solution is: 0.502281529540454
365,2290 solution is: 0.503235653963656
Break even at m = 2287
```

Approximate Analytical Solution

The only closed-form analytical "solution" that I've come across is an approximate analysis similar to that used to analyse lightning strikes (thanks to Iain Stewart for pointing this out). Q: In how many consecutive 'intervals' of time will lightning strike? If the strikes are independent the number of strikes per interval has a Poisson distribution with some parameter r. Recall that if N is a random variable with a Possion distribution, parameter r, then

$$P(N = n) = \frac{r^n e^{-r}}{n!}$$

Here, the intervals are days and the events are birthdays. Otherwise it's the same problem. So, the probability of *no* birthdays in an interval (day) is $P(N = 0) = e^{-r}$ where r is the rate per interval. The m we want (total number of people) is then $365 * r$ — think of this as r people per interval (day). Making the approximating assumption that the intervals are independent, the probability that *every* interval (day) has at least one birthday is $(1 - e^{-r})^{365}$. Setting this to 0.5, our required break-even point, r is approximately 6.267, making $m = 365 * 6.27 = 2287.59$. Note that the approximation is bad for small n (the number of days). For $n = 365$ the answer is amazingly close to the exact value!