The Birthday Problem

Note Title 18/10/2007

This is a traditional probabilistic problem: given n people, whose birthdays are uniformly distributed over the 365 days of the year (ignoring leap years), find

$$\mathbb{P}(\text{at least 2 people share a birthday}).$$

It is easier to note that

$$\mathbb{P}(\text{at least 2 share a birthday}) = 1 - p_n, \qquad (P1)$$

where

$$p_n = \mathbb{P}(\text{all } n \text{ people have different birthdays}).$$

Now

$$p_n = \frac{365 \cdot 364 \cdot 363 \cdot \cdots \cdot (365-n+1)}{365^n}$$

$$p_n = \left(1 - \frac{1}{365}\right)\left(1 - \frac{2}{365}\right) \cdots \left(1 - \left(\frac{n-1}{365}\right)\right). \qquad (P2)$$

It's now an easy matter to calculate $1-p_n$; here are some sample values:

n	$1-p_n$
15	0.15
23	0.51
35	0.99

This is the surprising part: given 23 people, there is an evens chance of a shared birthday.

Can we get some feel for the very rapid decrease of p_n? First take logarithms:

$$\ln p_n = \sum_{k=1}^{n-1} \ln\left(1 - \frac{k}{365}\right). \qquad (P2)$$

Now (see the end for a derivation)

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots, \qquad (P3)$$

and the series is convergent for $|x| < 1$. Thus

$$\ln(1-x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \frac{x^4}{4} - \cdots$$

$$\leq -x. \qquad (P4)$$

Hence

$$\ln p_n = -\sum_{k=1}^{n-1}\left(\frac{k}{365} + \frac{k^2}{2 \cdot 365} + \cdots\right)$$

$$\simeq -\frac{\sum_{k=1}^{n-1} k}{365} = -\frac{n(n-1)}{730} \qquad (P5)$$

and

$$\ln p_n \leq -\frac{n(n-1)}{730}. \qquad (P6)$$

Taking exponentials,

$$p_n \simeq e^{-n(n-1)/730} \qquad (P7)$$

and

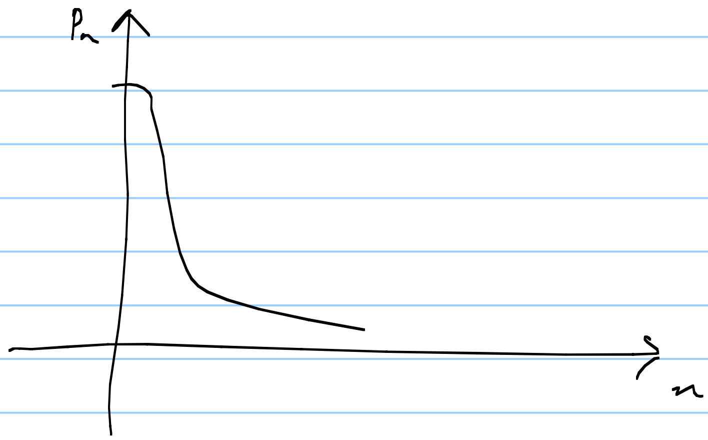
$$p_n \leq e^{-n(n-1)/730} \leq e^{-(n-1)^2/730}. \qquad (P8)$$

Thus $p_n = \frac{1}{2}$ implies

$$\ln \frac{1}{2} \simeq - \frac{(n-1)^2}{730},$$

or
$$(n-1)^2 \simeq 730 \ln 2$$
$$\simeq 730 * 0.6 = 438.$$

Since $21^2 = 441$, this implies $n-1 \simeq 21$, or $n = 22$, which is excellent agreement. However, it's more important that we have $p_n \leq e^{-(n-1)^2/730}$, because this tells that the coincident birthday probability decays like a Gaussian:



Now consider the closely related problem: suppose we have a pseudo-random number generator which is claimed to generate numbers uniformly in $\{1, 2, \ldots, N\}$ and we take n samples. On modern computers, N is a large power of 2. Then

$$p_n = \mathbb{P}(\text{all } n \text{ different}) = \prod_{k=1}^{n-1} \left(1 - \frac{k}{N}\right)$$

and, as before

$$\ln p_n = \sum_{k=1}^{n-1} \ln\left(1 - \frac{k}{N}\right) \leq \frac{-n(n-1)}{2N},$$

i.e. $\quad p_n \simeq e^{-n(n-1)/2N} \quad$ and $\quad p_n \leq e^{-n(n-1)/2N}$.

This gives us a simple test: does the pseudo-random generator produce numbers consistent with these estimates for p_n?