The Birthday Paradox

The "Birthday Paradox" refers to the fact that it is much more likely that two people in a large group will have the same birthday than it is that someone in that group will have a birthday on a *specific* day. For example, in order to have a 50% probability that one person in a group has a specific birthday requires a group of 253 people, but in a group of only 23 people there is a 50% probability that at least two of those people have the same birthday.

Of course, this is not really a paradox, just a result that is non-intuitive. The birthday paradox arises in a number of situations, including cryptography and the design of secure hashing algorithms, digital communications systems and, most recently, in the performance of the random shuffling feature on the Apple iPod. Many users of the iPod noticed that songs by the same group seemed to play unusually close together, leading to the speculation that Apple engineers did not implement a truly random shuffle algorithm. (Apparently some iPod users actually thought their iPod might have some sort of intelligence that "knew" what songs they wanted to hear.) See the following links for news stories on this:

http://www.nytimes.com/2004/08/26/technology/circuits/26ipod.html?ex=1251345600&en=a81c762718429d7e&ei=5088&partner=USERLAND

http://www.wired.com/news/culture/0,1284,68893,00.html?tw=wn_story_page_prev2

Here we will derive the key result for the birthday paradox. We will initially put the problem in the context of birthdays, but the results are much more general than this.

The question to consider is how many people must be in a room so that the probability that at least two people will have the same birthday is 0.5. Let N be the number of possible values, in this case 365. Let M be the number of people in the room. Let x_i be the birthday date of the i-th person in the room, for $i = 1, 2, 3, \ldots M$. We begin by finding the probability that no two birthdays of people in the room are equal.

First consider the case with only two people (x_1, x_2) in the room ($M = 2$). Then,

$$P\{\text{no two birthdays are equal}\} = P\{x_1 \neq x_2\} = 1 - P\{x_1 = x_2\} = 1 - \frac{1}{N}.$$

Now add a third person to the room, (x_1, x_2, x_3), for M=3:

$$
\begin{aligned}
P\{\text{no two birthdays are equal}\} &= P\{x_1 \neq x_2 \,(\text{for } M=2)\}\, P\{x_3 \neq x_1 \text{ and } x_3 \neq x_2, \text{ given that } x_1 \neq x_2\} \\
&= P\{x_1 \neq x_2 \,(\text{for } M=2)\}\left[1 - P\{x_1 = x_3 \text{ or } x_2 = x_3, \text{ given that } x_1 \neq x_2\}\right] \\
&= P\{x_1 \neq x_2 \,(\text{for } M=2)\}\left[1 - P\{x_1 = x_3\} - P\{x_2 = x_3\}\right] \\
&= \left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)
\end{aligned}
$$

This process continues, so that the probability that no two birthdays are equal for a room of M people is,

$$
\begin{aligned}
P\{\text{no two birthdays are equal}\} &= \left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\left(1-\frac{3}{N}\right)\ldots\left(1-\frac{M-1}{N}\right) \\
&= \prod_{i=1}^{M-1}\left(1-\frac{i}{N}\right)
\end{aligned}
$$

Now we need to make an approximation, which is generally good if N is large. For small x we know that $e^{-x} \simeq 1-x$, so for $N >> M$ we can write,

$$
\begin{aligned}
\prod_{i=1}^{M-1}\left(1-\frac{i}{N}\right) &\simeq \prod_{i=1}^{M-1} e^{-i/N} = e^{-1/N} e^{-2/N} e^{-3/N} \ldots e^{-(M-1)/N} \\
&= e^{-\left[1+2+3+\ldots(M-1)\right]/N} = e^{-\frac{M(M-1)}{2N}}
\end{aligned}
$$

If we choose this probability to be 50%, then we can solve for the group size:

$$0.5 = e^{-\frac{M(M-1)}{2N}},$$

$$2N\ln 2 = M(M-1),$$

$$M \simeq \sqrt{2N\ln 2} \simeq 1.17\sqrt{N}\,.$$

For the case of birthdays, with N = 365, this gives M = 23 people. For an iPod with, say, 100 albums, there is a 50% probability that you will hear at least two songs from the same album after playing only 12 songs. Of course, your mileage may vary unless the sample size is very large.