Matching Criminals - Nrich Live Question

Name: James Kim
Date: 23rd August 2016
School: West Island School

Prompt

This problem considers a surprising result that can have some real life implications. Read the text below, and then take a look at the questions that follow. DNA profiling is an invaluable tool for the police and the courts to identify criminals. This is because it's extremely accurate: the chance that a random person will match a DNA profile taken from a crime scene is estimated to be less than 1 in a billion.

However, when it comes to chance and probability, things aren't always as straightforward as they seem. You have to very careful to do your sums properly before jumping to conclusions.

As an example, let's take a controversy sparked by the discovery of a lab worker in the US state of Arizona in 2001. While she was sifting through a database containing DNA profiles of 65,000 convicted criminals the lab worker found two matching profiles. The two convicts to whom the profiles belonged were unrelated. Subsequent research showed that there were many more matching pairs in the database.

This surprised people, because the chance of one profile matching a given profile is so small. Some people suggested that something must be seriously wrong with the technology used to analyse and compare DNA samples. This would pose a terrible problem because if the technology gives rise to unnaturally many matches, then many innocent people might be wrongly convicted of crimes.

We can model a similar situation where the probability of any two random people matching is 1 in 225, and we have a database of 30 people. Each person is given a random number between 1 and 225; this represents their DNA profile.

Imagine you are one of the group of 30 people on the database. What is the probability that someone else has the same number as you?

To calculate the probability that somebody else has the same number as you, the total probability theorem can be applied.

By definition of the theorem, we know that

$$Pr[P] + Pr[P'] = 1$$

given that the property P is an element of some set S, and P' is its respective inverse.

In this case, let P be the probability that another person has the same number, let P' be the probability that all people have different numbers, and let S be the set of all possibilities.

We now know that

Pr[Another person has the same number] = 1 - Pr[All people have different numbers]

To get the Pr[All people have different numbers], we have to consider every single person. In this sense, we have to consider the probability that your number is different from all 29 other people and to do this, the multiplication rule can be used.

As given by the question, we know that the

$$\textit{Pr [two random people are the same]} = \frac{1}{225}$$
$$\textit{Pr [two random people are different]} = \frac{224}{225}$$

By the multiplication rule of probabilities, which states that if events A and B come from the same sample space S, the probability that both A and B occur is equal to the probability the event A occurs times the probability that B occurs, given that A has occurred, we can tell that

$$\text{Pr[All numbers are different]} = (\frac{224}{225})^{29} = 0.879$$

Hence

$$\text{Pr[Another person has the same number]} = 1 - (\frac{224}{225})^{29} = 1 - (0.879) = 0.121$$

How likely do you think it is that there will be at least one match amongst the 30 people in the database?

Although by intuition, it may seem that there is a low probability that there does not exist a matching pair of numbers due to the initial probability of 1/225, but in reality, this is not the case. This misinterpretation is mainly due to how people tend to comprehend probability in linear terms instead of exponentials. This principle is clearly demonstrated though the Birthday Paradox.

What is this probability?

To calculate the probability that there exists **at least** a pair of people with the same number in a given group, the total probability theorem can be applied similarly to the previous question.

In this case, let P be the probability that there exists a pair of 2 or more people in set S, and let P' be the probability that all people in set S have different numbers.

Hence we now know that

Pr[Exists a pair with the same number] = 1 - Pr[All numbers are different]

To get the Pr[All numbers are different], we have to consider every single pairing and apply the multiplication rule again. In this sense, the cardinality of pair combinations possible is

$$\binom{30}{2} = (\frac{30!}{28!2!}) = 435$$

As given by the question, we know that the

Pr [two random people are the same] $= \frac{1}{225}$
Pr [two random people are different= $\frac{224}{225}$

Pr[All numbers are different] $= (\frac{224}{225})^{435} = 0.144$

Hence Pr[Exists a pair with the same number] $= 1 - 0.144 = 0.856$

Why is it so much more likely that two people will share the same number than someone sharing your number?

It is much more likely that there exists a pair of two people who share the same number, because this inquiry requires **all** numbers to be different, whereas the probability that someone shares **your number** only requires the other people to be different from *only you.*

Does this help to explain why so many pairs were found in the Arizona database?

Yes, this explains why there were many pairs found in the Arizona Database. If we were to implement the same method of calculation from Part 4, it would be as below.

Pr[Exists a pair with the same DNA] = 1 - Pr[All DNA are different]

To get the Pr[All DNA are different], we have to consider every single pairing and apply the multiplication rule again. In this sense, the cardinality of pair combinations possible is

$$\binom{65000}{2} = (\frac{65000!}{64998!2!}) = 2,112,467,500$$

As given by the question, we know that the

Pr [two random people are the same] $= \frac{1}{10^9}$
Pr [two random people are different]$= \frac{999,999,999}{10^9}$

$$\text{Pr[All DNA are different]} = (\frac{999,999,999}{10^9})^{2,112,467,500} \approx 0.121$$

Hence Pr[Exists a pair with the same DNA] $= 1 - 0.121 = 0.879$