
DNA evidence and databases

A criminal leaves DNA at a crime-scene. Later on, Joe is charged with the crime, with the main evidence being that Joe's DNA fingerprint matches the fingerprint from the crime scene. The probability of such a match between two unrelated people is $m = 10^{-6}$. How should this evidence be used? Does it matter whether Joe was picked up by trawling through a database of DNA profiles? *Does the size of the database matter?*

Let's study this question in the cleanest way possible, introducing simplifying assumptions so we can focus on the key issue. We assume that the number of people in the city who could have committed the crime is N_C (say $N_C = 10^5$), that there is no other evidence apart from the DNA evidence, and that all the N_C suspects have equal prior probability of being the criminal. We assume that a random subset of the population of size N_D are fingerprinted. (D is mnemonic for database.) Possible values for N_D include 1 – the police stop one random individual, Joe, and test him alone – and any number up to N_C . In a plausible story, we could say $N_D = 10^4$ – ten percent of the population are fingerprinted.

Now, among these N_D , we assume that there is exactly one match (Joe). We assume that the $(N_C - 1)$ innocent persons' DNA fingerprints are generated independently and identically with a probability $m = 10^{-6}$ of matching the fingerprint found at the scene. So the total number of innocent people in the city who match, k , has a Poisson distribution

$$P(k | m, N_C) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad (51.1)$$

where $\lambda = m(N_C - 1)$ is the expected number of innocent matches. Assuming N_C is large, we replace $\lambda = m(N_C - 1)$ by $\lambda = mN_C$ from here on. The actual total number of matches in the city, innocent and guilty, is $k + 1$. Now, if we knew k , then the probability that Joe is the guilty party would be

$$P(\text{Joe is guilty} | k) = \frac{1}{k + 1}. \quad (51.2)$$

According to Tijms (2004)¹, the probability that Joe is guilty is given by averaging $1/(k + 1)$ under the Poisson distribution (51.1):

$$P_{\text{Tijms}}(\text{Joe guilty} | \text{Data}) = \sum_k P(\text{Joe is guilty} | k) P(k | m, N_C) \quad (51.3)$$

$$= \sum_k \frac{1}{k + 1} e^{-\lambda} \frac{\lambda^k}{k!} \quad (51.4)$$

$$= \frac{1 - e^{-\lambda}}{\lambda} \simeq 1 - \frac{\lambda}{2} + \dots \quad (51.5)$$

¹Henk Tijms (2004): Understanding Probability, Cambridge University Press.

For the numbers given above ($m = 10^{-6}$, $N_C = 10^5$), we find

$$P_{\text{Tijms}}(\text{Joe is guilty} \mid \text{Data}) = 0.95. \quad (51.6)$$

This answer is in the right ballpark, but it is incorrect.

Before we did any database searching, the prior distribution of k was the Poisson distribution (51.1). But the fact that a database search among N_D randomly chosen people generated one match gives information about the total number of innocent matches k . The correct distribution to average over is not the prior on k but the posterior.

Rather than working out the general answer for a database of size N_D , let's work out some instructive special cases. If *everyone* is in the database ($N_D = N_C$) and the number of matches found is 1, then the posterior distribution of k is a spike on $k = 0$; we are *certain* that Joe is guilty. All other suspects have been eliminated. (Our simplified model omitted other possibilities such as laboratory error and police fraud.)

If only one person is tested ($N_D = 1$), then the posterior distribution of k , given that the number of matches d found in the database is 1, is

$$P(k \mid d=1, N_D=1) = \frac{P(d=1 \mid k, N_D=1)P(k)}{P(d=1, N_D=1)}. \quad (51.7)$$

The likelihood function, assuming the subjects for the database are chosen at random from the population, is

$$P(d=1 \mid k, N_D=1) = \frac{k+1}{N_C}. \quad (51.8)$$

The correct answer in this case is thus:

$$\begin{aligned} P(\text{Joe is guilty} \mid \text{Data}, N_D=1, d=1) \\ = \sum_k P(\text{Joe is guilty} \mid k)P(k \mid m, N_C, N_D=1, d=1) \end{aligned} \quad (51.9)$$

$$= \frac{\sum_k \frac{1}{k+1} e^{-\lambda} \frac{\lambda^k}{k!} \frac{k+1}{N_C}}{\sum_k e^{-\lambda} \frac{\lambda^k}{k!} \frac{k+1}{N_C}} \quad (51.10)$$

$$= \frac{1}{\lambda+1} \simeq 1 - \lambda + \dots, \quad (51.11)$$

which for the numbers above is

$$P(\text{Joe is guilty} \mid \text{Data}, N_D=1, d=1) = \frac{1}{1.1} = 0.91. \quad (51.12)$$

So the probability that Joe is innocent is 0.09, roughly a factor of two bigger than the 0.05 given by Tijms (51.5, 51.6).

So the size of the database trawled through *does* matter.

The case $N_D = 1$ is rather implausible – the police stumble into a matching suspect immediately? The chance of that occurrence is the normalizing constant of our inference (51.7), which is $(\lambda + 1)/N_C$.

In practice, the number of suspects who are fingerprinted, N_D , will be larger. To compute the correct weight of evidence in court, the jury must be told the size of the database, and the result of all the comparisons with the database. If the number of matches found in the database were $d = 2$, our inferences about Joe's guilt would be very different. If $d = 2$ then the probability that Joe is innocent is greater than 1/2, independent of λ .