

STATISTICAL ISSUES IN DNA IDENTIFICATION AND TESTING

ŚAUNAK SEN*

UNIVERSITY OF CALIFORNIA SAN FRANCISCO

sen@biostat.ucsf.edu

INTRODUCTION

Advances in DNA technologies have dramatically altered forensic science, as well as medical testing. Although the objective of DNA forensic analysis (“DNA fingerprinting”) is very different from that of genetic testing, there are some statistical similarities. To explore these issues we will first review the basics of diagnostic tests from a statistical perspective. We will then apply those concepts in the context of DNA fingerprinting and genetic testing.

ANATOMY OF A DIAGNOSTIC TEST

Consider a test for the presence or absence of a disease. Denote the outcome of the test (“positive”, + or “negative”, -) as T , and the true state of the sample (“diseased”, + or “normal”, -) by D . The objective of the test result, T , is to shed light on the unknown state of nature, D . The outcomes can be represented in the following table:

		Test	
		Positive	Negative
Sample	Diseased	True	False
	Normal	False	True

The probability of a true positive for a diseased sample, $P(T=+|D=+)$, is the *sensitivity* of the test, while the probability of a true negative for a normal sample, $P(T=-|D=-)$, is the *specificity* of the test. However, most often we are really interested in is, the probability that the tested sample is diseased given that the test is positive (*positive predictive value*) and the probability that the tested sample is normal given a negative test result (*negative predictive value*).

To see the difference between the two approaches, consider the case of a disease present in 1% of the population for which we have a test that is 95% sensitive and 99% specific. If we apply the test to 10,000 individuals we would get the following expected number of outcomes.

		Test		Total
		Positive	Negative	
Sample	Diseased	95	5	100
	Normal	99	9801	9900
	Total	194	9806	10000

In this example, the positive predictive value is a modest 0.49, while the negative predictive value is about 0.9995. Thus the test is very good at ruling out the possibility of a diseased sample, and not so good at predicting if a sample is diseased. This is in spite of the high sensitivity and specificity of the test.

The principal reason for this behavior is the low *a priori* probability of a diseased sample. In fact, if the disease is present in 10% of the population, the picture changes dramatically.

* © 2006 Śaunak Sen; Last updated May 09, 2006.

Sample		Test		Total
		Positive	Negative	
Diseased		950	50	1000
Normal		90	8910	9000
Total		1040	8960	10000

The positive predictive value is now 0.91 and the negative predictive value is now 0.9944. The negative predictive value does not change much, but the positive predictive value changes dramatically. Thus, the predictive values are affected not only by the sensitivity and specificity, but also by the population prevalence.

Bayes Theorem

More generally, we can write the positive predictive value in terms of the corresponding odds ratio as

$$\frac{P(D=+|T=+)}{P(D=-|T=+)} = \frac{P(D=+)}{P(D=-)} \times \frac{P(T=+|D=+)}{P(T=+|D=-)}$$

Similarly we can write the odds for the negative predictive value as

$$\frac{P(D=-|T=-)}{P(D=+|T=-)} = \frac{P(D=-)}{P(D=+)} \times \frac{P(T=-|D=-)}{P(T=-|D=+)}$$

We can rephrase this by saying that the relationship between the predictive values, population prevalence, sensitivity and specificity can be summarized by Bayes theorem as follows:

$$\text{Posterior odds} = \text{Prior odds} \times \text{Likelihood ratio}$$

This equation shows how we can update our current knowledge using observed data by making explicit the contribution of *a priori* knowledge, and evidence from gleaned from the data alone.

This means that the result of a diagnostic test has to be interpreted in the context of the sample. More specifically, if the sample is from a high risk population, the positive predictive value is increased relative to a sample from a low risk population.

DNA IDENTIFICATION

Consider the forensic problem of trying to place a suspect onto a crime scene where biological material such as blood, semen, or hair has been found. Typically, the sample on the crime scene, as well as one from the suspect are typed at a few highly polymorphic markers.

Match probability

The test result is a “match” or “no match” based on the typing a certain number of common markers in both the sample from the crime scene and a suspect. There are two possibilities for truth, the crime scene sample is from the suspect or not. The commonly-reported number is the “match probability” which is the probability that a randomly-drawn person from the population will have a matching profile at the typed markers.

Let us consider the simple case of when a single marker is typed. Suppose the alleles A and B are observed, with frequencies p_A and p_B respectively in the database population. Then the probability of a match (or getting a person with genotype AB) is $2p_A p_B$. It is easily seen that rarer the observed alleles, the smaller the match probability. However this ignores the possibility of population structure in the database population.

It can be shown that in the presence of structure the match probability is[†]

$$2 \frac{(F + (1-F)p_A) (F + (1-F)p_B)}{(1 + F)(1 + 2F)},$$

where F is a measure of the extent of population structure (analogous to the measure F_{ST}).

To get an idea how population structure may affect the match probability, consider the case of two alleles, each with a 1% frequency in the population. The usual calculation of the match probability gives 0.0002. But increasing amounts of population structure undermine the strength of the evidence. In observed populations the value of F may range from 0.01 to 0.10, but this depends on the context.

Allele freq	Match Probability	Multiplication factor				
		F				
		0.01	0.05	0.10	0.25	0.50
0.10	0.02	1.2	1.8	2.7	5.6	10.1
0.05	0.005	1.4	3.3	6.4	17.6	36.8
0.01	0.0002	3.8	30.7	90.0	353.6	850.1
0.001	0.000002	117.2	2247.5	7712.7	33533.6	83500.1

In context: Bayes Theorem

Thus the match probability has to be interpreted in the context of the population from which the suspect is drawn from. It is also essential that we combine the evidence from the DNA test to other evidence using Bayes Theorem. Drawing an analogy from the disease testing problem, there are two states of nature – the crime sample is from the suspect, and that it is not. The match probability is the probability that we find a matching DNA profile from a randomly-drawn person (the complement of the specificity). The probability of matching given that the crime scene sample is from the suspect is close to 1 (the sensitivity). Thus the odds that the crime sample is from the suspect is

$$\text{prior odds} \times \frac{1}{p},$$

where p is the match probability. The prior odds depend on other non-DNA evidence about the crime and the suspect's involvement.

However, it is important to remember that “absence of evidence is not evidence of absence,” that is, other evidence may implicate a suspect even though the DNA evidence may not be convincing.

[†]Balding DJ and Nichols RA (1994) DNA profile match probability calculations: how to allow for population stratification, relatedness, database selection and single bands, *Forensic Science International*, 64:125-140.