1

Introduction

1.1 Statistics, forensic science and the law

Statistics has been playing an important role in forensic science and law. This is very natural, since statistics is the science in dealing with variability and uncertainty, which commonly arise in these two disciplines. In forensic science, data are collected from the crime scene or elsewhere, and statistics is often used to analyze these data. Scientists explain the statistical findings and provide their interpretations to various concerned parties including the client, jury, lawyer and judge. Recently, several books were published on the use of statistics in forensic science and in the courtroom (Aitken and Taroni 2004; Gastwirth 2000; Good 2001; Lucy 2005).

According to (Lucy 2005, p3), a brief inspection of the *Journal of Forensic Sciences* for the years between 1999 and 2002 indicates that about half of the articles have some kind of statistical content. It is noticed that the sort of statistical methods used can vary from the elementary tools such as percentages, means and standard deviations to the more sophisticated techniques including tests of statistical hypotheses, regression and calibration, and classification. An update in glancing through the articles in the *Journal of Forensic Sciences* for the years of 2005 and 2006 indicates that the phenomenon persists, i.e. about half of the articles have some kind of statistical content. Besides those statistical methods mentioned by Lucy (2005), we also find other more complex methods, such as cluster analysis, logistic regression and Fisher's exact test. Moreover, we also notice that about a quarter of the articles are on DNA profiling. Nearly all these DNA articles involve some kind of statistical analyses, ranging from elementary statistical methods to more complicated techniques such as the least-square deconvolution.

1.2 The use of statistics in forensic DNA

DNA profiling has become one of the most commonly used techniques for human identification since its introduction by Jeffreys *et al.* (1985). It is one of the most important tools in forensic science. Nowadays, many forensic laboratories including the Hong Kong Government

Statistical DNA Forensics: Theory, Methods and Computation Wing Kam Fung and Yue-Qing Hu

Laboratory have the largest teams of scientists working in DNA forensics. DNA can be found in blood, hair/hair root, bone, semen and body fluid such as saliva and sweat. No two persons, except for identical twins, have the same DNA sequence. The current DNA profiling technology uses only a number of genetic markers, and so a unique identification may not be assured. Nevertheless, the technique is widely employed and accepted in courtrooms due to its highly discriminating power and reliability. The US National Research Council (NRC) released two reports on the use of the technique in 1992 and 1996. In NRC II (National Research Council 1996), many discussions were provided on the statistical issues of forensic DNA, and several recommendations related to the proper use of statistics were given. Since NRC II, a few books on the use of statistics in DNA forensics have been published (Balding 2005; Buckleton *et al.* 2005; Evett and Weir 1998).

DNA profiling is not only commonly used in forensic investigation, but also leads to a lot of research in this area. Nowadays, this kind of research constitutes the highest percentage of articles in respectable forensic science journals. In 2007, a daughter journal of Forensic Science International (FSI), *Forensic Science International: Genetics* (*FSI Genetics*), has been newly launched. According to the announcement in the founding issue of *FSI Genetics*, 46% of submissions to *FSI Genetics* fall in the area of forensic genetics, indicating that this discipline can readily support its own journal.

The following quote is taken from the founding volume of *FSI Genetics (2007)*:

> Although forensic genetics is a discipline a century old (the discovery of the ABO group by Karl Landsteiner can be considered the birth of this field), the introduction of DNA profiling to forensic analysis following the development of this technique by Alec Jeffreys and co-workers, 20 years ago, has had a tremendous impact on forensic genetics. The amount of work in this field has increased enormously since 1985, with an increasing number of papers published in this area. This increase shows no signs of slowing down with many new technologies and applications being reported. Major advances in molecular biology and computer technology—allowing DNA samples to be obtained from ever smaller quantities of biological material—are continuously being reported along with new and exciting applications of DNA technology to the analysis of non-human material (crime scene analysis, tracking the illegal trade in endangered species and bioterrorism), or the building and appropriate management of DNA databases is expanding outside of the traditional areas of criminal investigation.
>
> Forensic genetics is now a reasonably mature field, and generates sufficient high quality content to support a dedicated journal.
>
> The scope of the new journal would include most of the forensic genetics topics such as (among others):
>
> - Biostatistical methods in forensic genetics.
>   - Evaluation of DNA evidence in forensic problems (such as paternity or immigration cases, criminal casework, identification), classical and new statistical approaches.

In fact, a high proportion of the papers in forensic genetics has used some sort of statistics. In many situations, simple statistics such as the percentage, mean and standard deviation are

sufficient, while in some others, more advanced statistical analyses are needed. The following two articles selected from *FSI 2006* indicate the sorts of advanced techniques used.

Shepard and Herrera (2006) studied allelic frequencies of 15 STR loci from 150 unrelated persons from an Iranian population. Common statistical measures such as the gene diversity index, power of discrimination, power of exclusion and tests for Hardy–Weinberg equilibrium, etc. were constructed. The more advanced statistical techniques–phylogenetic analysis with neighbor-joining trees and multi-dimensional scaling analysis–were performed using F_{st} measures generated from 13 worldwide, geographically targeted populations. Bonferroni adjustment for multiple comparisons and statistical bootstrap analysis were also conducted. Based on the statistical findings, the authors discussed the appropriate choice of databases on which to base forensic calculations for populations located in geographic intersections.

Hammer *et al.* (2006) considered a set of 61 Y-SNPs for a sample of 2517 individuals from 38 populations to infer the geographic origins of Y chromosomes in the United States and to test for paternal admixture. Sophisticated statistical techniques, including hierarchical genetic structuring based on an analysis of molecular variance and multi-dimensional scaling for clustering, were chosen. From the statistical findings, it was inferred that both inter-ethnic admixture and population subdivision might contribute to fine scale Y-STR heterogeneity within US ethnic groups.

Why has statistics attracted more attention in DNA forensics than in other areas of forensic science? Fung *et al.* (2006) have summarized the following, among other possible reasons:

> First, DNA profiling is generally scientifically unambiguous and very powerful. Since the DNA evidence is repeatable, statistical evaluation would then be possible and in most situations objective.
>
> Second, when there is a match to the DNA evidence, people would like to know how likely there is a random match.
>
> Third, extremely small probabilities are commonly encountered in DNA profiling, and people are curious about their derivations and interpretations (note: these probabilities are sometimes interpreted incorrectly, e.g. prosecutor's fallacy).
>
> Fourth, many forensic scientists are not that familiar with statistics, particularly on different approaches of the subject.
>
> Fifth, some problems such as kinship determinations and DNA mixtures need complex statistical analysis.

It is the fifth point about kinship determinations and assessment of DNA mixtures that requires complex statistical analysis; a major aim of this book is to provide details on the statistical treatment of such problems. In doing so, the other points will also be touched upon.

1.3 Genetic basis of DNA profiling and typing technology

1.3.1 Genetic basis

NRC I and NRC II (National Research Council 1992, 1996) give comprehensive accounts of the general principles of DNA profiling. The following paragraphs on the genetic basis of DNA typing come from Chapter 2 of NRC II.

In higher organisms, the genetic material is organized into microscopic structures called *chromosomes*. A fertilized human egg has 46 chromosomes (23 pairs). A chromosome is a very thin thread of DNA, surrounded by other materials, mainly protein. The DNA thread is actually double – two strands coiled around each other like a twisted rope ladder with stiff wooden steps. The basic chemical unit of DNA is the nucleotide, consisting of a base. There are four kinds of bases, designated A, G, T and C. The total DNA in a genome amounts to about 3 billion nucleotide pairs. A gene is a segment of DNA, ranging from a few thousand to more than a hundred thousand nucleotide pairs. The position on the chromosome at which a particular gene resides is its *locus*.

Alternative forms of a gene are called *alleles*. If the same allele is present in both chromosomes of a pair, then the person is *homozygous*; if the two alleles are different, then the person is *heterozygous*. A person's genetic makeup is the *genotype*. Genotype can refer to a single gene locus with two alleles, A and a, in which case the three possible genotypes are AA Aa, and aa; or it can be extended to several loci or even to the entire set of genes. In forensic analysis, the genotype for the group of analyzed loci is called the DNA *profile*.

1.3.2 Typing technology

Currently, short tandem repeat (STR) loci are most commonly used for DNA profiling. Usually, about 10 or more unlinked autosomal (the 22 pairs of chromosomes, but not the sex chromosomes) loci are used in practice. After some laboratory procedures, including DNA extraction and the polymerase chain reaction (PCR) process, the STR profiles can then be obtained. For more details on the STR typing technology, interested readers can refer to (Buckleton *et al.* 2005, Chapter 1) and (Balding 2005, Chapter 4).

Figure 1.1 shows the STR profile of a DNA sample from a crime scene obtained by ABI machines and software. Only the three autosomal loci of the yellow panel are shown for illustration. The upper panel corresponds to the allelic ladders of the standard markers. From the figure, we notice that the genotypes at the three loci are respectively 7/11 at locus

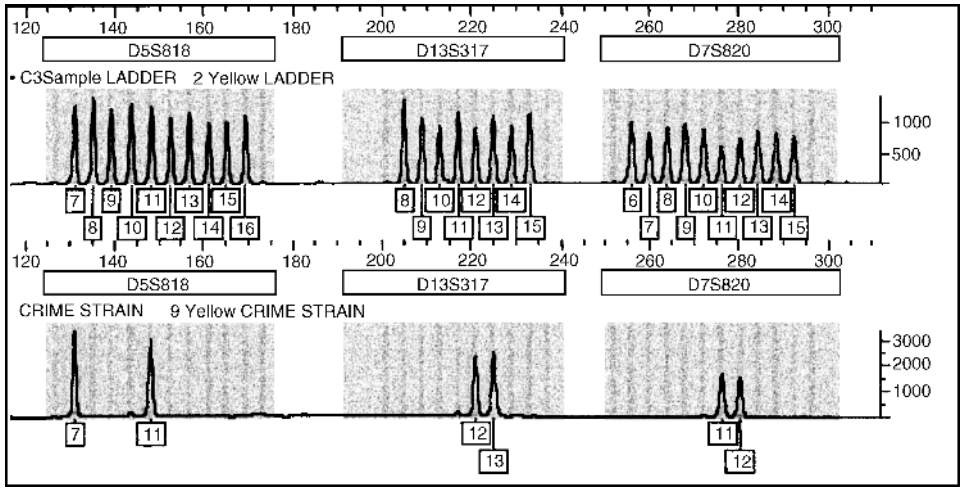

Figure 1.1 An STR profile of a DNA sample from a crime scene, obtained by ABI machines and software.

D5S818 (of chromosome 5), 12/13 at D13S317 (of chromosome 13) and 11/12 at D7S820 (of chromosome 7). The values 7 and 11 at D5S818 may be called allele sizes, which represent the numbers of repeat DNA units in the two alleles. The STR locus has the property that its allele size is discrete and so is easy to interpret and has little ambiguity. The commonly used STR loci usually have slightly below 10 to over 20 alleles, giving a large number of possible genotypes at each locus. In the Hong Kong Chinese population database (Wong *et al.* 2001), it just happens that there are eight different alleles for each of the loci mentioned above, resulting in $8 \times 9/2 = 36$ possible genotypes at each locus.

To show the discriminating power of DNA profiling, we assess the frequency of the DNA profile in Figure 1.1 in the Hong Kong Chinese population. According to Wong *et al.* (2001), the allele frequencies are $p_7 = 0.035$ and $p_{11} = 0.252$ at D5S818; $p_{12} = 0.099$ and $p_{13} = 0.023$ at D13S317; and $p_{11} = 0.376$ and $p_{12} = 0.230$ at D7S820. The frequency of the DNA profile at all three loci may be evaluated as $(2 \times 0.035 \times 0.252) \times (2 \times 0.099 \times 0.023) \times (2 \times 0.376 \times 0.230) = 1.39 \times 10^{-5}$ under Hardy–Weinberg and linkage equilibria (discussion on their validity is given in Chapter 3). In other words, about 1 in 72 000 $[= 1/(1.39 \times 10^{-5})]$ persons in the local Chinese population has such a DNA profile. This shows the highly discriminating power of the technique if a suspect is arrested and his/her genotype is found to match with the crime stain profile.

1.4 About the book

This book aims to introduce the basic statistical theory and methods for the evaluation of DNA evidence. Readers are assumed to have little background knowledge in statistics and probability. Thus, we start by considering simple cases first and then proceed to analyze more complex problems. We illustrate with many examples, so that readers can not only grasp the basic concepts, but also understand the more advanced analyses. The book covers three main applications of DNA profiling, namely identity testing, determination of parentage and kinship, and interpretation of mixed DNA stains. Moreover, we place emphasis on the computational aspects of statistical DNA forensics. Computer programs are available at `http://www.hku.hk/statistics/EasyDNA/` for possible use. Readers can use the software to check the numerical findings of the examples given in the book. This can help readers to understand and appreciate the theory and methods behind statistical forensic DNA analysis.

The remainder of the book is organized as follows. Chapter 2 provides the basic probability and statistics that are commonly used in later chapters. Chapter 3 discusses fundamental concepts and introduces some statistical measures in population genetics. The statistical evaluation of single source samples or identity testing, including the theory of subpopulation models and the problems involving relatives, is studied thoroughly. The common parentage identifications are discussed in Chapter 4, while the complex kinship determinations are considered in Chapter 5. The associated computer software can provide a convenient means to analyze those particular paternity and kinship problems. Although the methods and software are illustrated with STR profiles, they can also be applied to analyze single-nucleotide polymorphism (SNP) profiles. Chapters 6 and 7 are on the statistical interpretation of DNA mixture. The associated formulas are often complicated and so the more technical derivations are put in the last section of each chapter. Thus, the reader can focus on the application of the calculating formulas in practical problem without being distracted by the technical derivations. The last chapter (Chapter 8) discusses some other issues in statistical DNA forensics, such as the Y-STR marker, peak information and database search, etc.