Statistical Science
1991, Vol. 6, No. 2, 175–205

Inferences Using DNA Profiling in Forensic Identification and Paternity Cases

Donald A. Berry

Abstract. Forensic laboratories use lengths of fragments from several locations of human DNA to decide whether a sample of body fluid left at the scene of a crime came from a suspect or whether a sample recovered from a suspect's clothing is the victim's. Using an inferential approach called "match/binning," they first decide whether there is a match between the lengths of DNA fragments from the suspect and crime samples. If there is a match, they then calculate a "match proportion." This is the proportion of a data base of DNA fragment lengths that would similarly match, that is, occur in an interval or "bin" containing the fragment length of the crime sample.

Match/binning is a reasonable inferential method in a scientific setting, and in other settings that allow for flexibility, but it has several characteristics that make it undesirable for use in courts. One is that it is based on a yes/no decision: there is an arbitrary cut-off point and some fragments deemed not to match can be arbitrarily close to others that do match. Another is that the same match proportion applies for suspects whose fragment lengths just barely match the lengths of the corresponding fragments in a crime sample as for suspects whose fragment lengths match perfectly.

This article describes an alternative approach, one that is not based on a yes/no match criterion. The distribution of a laboratory's measurement errors is used to infer the form of the likelihood function. Then the likelihood ratio of guilt to innocence is calculated and Bayes' theorem is applied. The focus of this approach is the contribution of the DNA evidence to the probability that a suspect is guilty. An important step is estimating the population distribution of fragment lengths, attempting to account for both laboratory measurement error and sampling variability.

The two approaches are compared in an actual murder case (*New York v. Castro*). Applying a laboratory's match criterion literally resulted in an exclusion, but its scientists claimed a match and calculated a match proportion that was very small. Applying Bayes' theorem shows that the correct conclusion is far less clear.

DNA profiling is also useful in inferring parentage, for example in cases of disputed paternity. Bayes' theorem allows for calculating the probability that an alleged father is the true father.

Key words and phrases: DNA fingerprinting, DNA sequence weights, restriction fragment length polymorphisms, multiple DNA probes, probability of guilt, probability of paternity, forensic identification, likelihood ratio, Bayes' theorem, Bayes factor, reference populations, measurement errors, lognormal errors, sampling variability, density estimation, normal kernels.

1. INTRODUCTION

Police investigators have a forensic tool with enormous potential for solving some types of crimes: DNA profiling (Jeffreys, Wilson and Thein, 1985a, b). (I will avoid the commonly used term "DNA fingerprinting" because the fingerprinting analogy

Donald A. Berry is Professor, School of Statistics, University of Minnesota, Minneapolis, Minnesota 55455, and Professor, Institute of Statistics and Decision Sciences and Comprehensive Cancer Center, Duke University, Durham, North Carolina 27706.

is not perfectly apt: identical twins have identical DNA but different fingerprints.) Such crimes include rapes in which the rapist's semen can be recovered from the victim or from the crime scene. They also include cases in which the criminal leaves his blood or other tissue at the scene of the crime or gets a victim's blood on his clothes or other possessions.

Forensic laboratories compare the molecular weights of fragments of DNA from the suspect sample and crime sample and decide whether the two fragments could have come from the same individual. Deciding whether there is a match is complicated by measurement error and laboratory process variability. Inferring whether two samples came from the same individual using DNA profiling is qualitatively the same as when using other genetic characteristics, such as ABO blood type. But when laboratories use hypervariable loci (locations on the genome where DNA fragment lengths tend to differ greatly among individuals), the number of possible molecular weights is enormous, so DNA profiling is potentially much more powerful than using genetic systems that have a small or moderate number of phenotypes.

The purpose of this article is to consider the inferential process of deciding whether a suspect is guilty. I will first describe and criticize the way in which DNA profiling laboratories currently make inferences and then show how to improve it. My focus is the assessment of the conditional probability of guilt given of DNA evidence. Some of the calculations I suggest are novel, but the overall approach is not new; of special significance are papers by Lindley (1977); Evett, Cage and Aitken (1987); Gjertson, Mickey, Hopfield, Takenouchi and Terasaki (1988); and Werrett, Gill, Evett, Lygo and Sullivan (1989). In Section 8, I will address the use of DNA profiling in cases of disputed paternity. These are usually civil cases, but they can be criminal cases if a rapist may have fathered the victim's child.

The FBI's forensic laboratories are the most important DNA profiling labs in the United States, even though they were rather late getting involved. The main commercial forensic laboratories that do DNA profiling in the United States are Lifecodes Corporation of Valhalla, New York, and Cellmark Diagnostics of Germantown, Maryland. All three use electrophoresis and restriction fragment length polymorphism (RFLP) analysis. Restriction enzymes are used to cut the DNA at specific sequences of DNA. The resulting fragments vary in size among individuals, but all resulting fragments from the same site on the genome of the same individual are identical; in particular, they have the same molecular weight.

Electrophoresis involves giving DNA samples an electrical charge and placing them in a gel. Technicians set up an electrical field in the gel to move DNA fragments distances that depend on their weight. The DNA is then transferred onto a nylon membrane and the two strands that make up the double helix are separated. Radioactive DNA probes that have been designed to attach themselves to particular fragments of DNA are added. The locations in the gel of these fragments are indicated as bands on X-ray film, called an *autoradiogram* and allow technicians to measure the distances traveled. These distances are converted to molecular weights or "band weights" (in kilobases (kb): thousands of Watson-Crick base pairs).

Single-locus probes yield most easily to quantification and are being used increasingly by forensic laboratories. A single-locus probe selects fragments of DNA from a single site on the genome. So a single-locus probe shows two bands, one paternal and the other maternal. Using several single-locus probes increases discrimination power. Multilocus probes can show any number of bands, depending on the number of sites selected by the probe.

To have a particular setting, suppose a criminal spills his blood at the scene of the crime. A suspect is identified and his DNA fragment lengths are compared with those of the crime sample. Inferential problems in other settings are similar (except that cases of disputed parentage—see Section 8—are qualitatively different). Perfect identification is not possible because other people may have DNA fragments that weigh the same as the crime sample, and even if two fragments have different molecular weights, they may not be distinguishable because of laboratory measurement error. The problem of inference is to take account of these possibilities, and to account for sampling variability, and decide whether the suspect is guilty. The ultimate question, of course, is the purview of the jury (or judge); the forensic statistician's role is to show the jury how to incorporate the DNA evidence into the answer.

2. QUANTIFYING INFERENCE: CURRENT PRACTICE

For each probe, technicians compare the suspect's band weights with those of the crime sample. Scientists at Lifecodes Corporation (Baird et al., 1986) indicate that a match occurs if there is no "detectable difference" between the band weights of the suspect and crime sample. And some other laboratories use such an informal "visual-match" criterion. If they claim a match, they estimate a "match proportion." This is the proportion of some reference population that would similarly match. For this purpose they use a data base of estimated band weights (for each probe individually).

Calculating a proportion of matches requires a formal match criterion. It is essential that the same criterion be used to decide whether the suspect matches as is used to decide whether someone who is not a suspect matches. Obviously, laboratories do not compare everyone in their data base to decide who are "visual matches." Several courts—including the Minnesota Supreme Court in *Hennepin County v. Schwartz*—have reacted negatively when a band weight of the suspect did not meet the match criteria that was used to calculate a match proportion.

Lifecodes and Cellmark call a band weight in the data base a match if the distance between it and the crime sample's band weight is less than k standard deviations of the measurement error. (This measurement error refers to the difference between two band weights and so the s.d. is $\sqrt{2}$ times that of a single measurement.) The FBI's procedure is based on preset bins; they take the proportion of the reference data base that falls in the bin containing the crime sample as the match proportion. The widths of the FBI's bins vary, but they are generally larger than the $\pm k$ s.d. bins of Lifecodes and Cellmark.

Baird et al. (1986) indicate that Lifecodes once used $k = 2$. Balazs, Baird, Clyne and Meade (1989) mentions $k = 3$. Morris, Sanda and Glassberg (1989) describe a matching procedure once used by Lifecodes; Lander (1989b) shows that it is approximately equivalent to using $k = 2/3$. As of July 1989, Lifecodes used $k = 3$ (M. Baird, personal correspondence). Evett, Werrett, Gill and Buckleton (1991) point out that any such cutoff is arbitrary: Why should 2.95 s.d.'s be a match and 3.05 s.d.'s be exclusionary? And shouldn't a perfect match be stronger evidence than one in which the discrepancy is 2.95 s.d.'s?

If the criminal has a *true* band weight that the suspect does not have, then the two individuals have to be different. So a suspect is excluded if any band weight does not match the crime sample's. Match/binning procedures exclude a percentage of guilty individuals because of measurement errors. And this incorrect exclusion rate increases with the number of probes used. Assuming independent normal measurement errors and four single-locus probes (and therefore, typically eight bands), $k = 2/3$ excludes about 99% of guilty individuals; $k = 1$, 2 and 3 exclude about 95%, 31% and 2%, respectively.

Based on 70 duplicate measurements, Baird et al. (1986) estimate the standard deviation for the difference between two measurements of a band weight to be 0.6% of the band weight; this corresponds to $0.6\%/\sqrt{2}$, or about 0.42% for a *single* measurement of a band weight. Baird provided me with several sets of replicate measurements that make it clear that 0.6% is too small and that 0.85% is better; this corresponds to $0.85\%/\sqrt{2}$, or about 0.6% for the s.d. of a single measurement. The assumption that s.d. is proportional to band weight is roughly supported by the data, although the fit is not perfect and the data are not comprehensive. The measurement s.d. is of critical importance for *all* inferential procedures that have been proposed. More and better-designed experiments are necessary before using an estimate of s.d. in so serious a setting as a murder trial.

The match proportion for a single band of a single-locus probe is typically small, especially for hypervariable loci and small k. In the heterozygous case there are two bands, although if their weights are sufficiently close together, the two bands may appear to be one. (Resolving power depends on the quality of laboratory equipment; according to Peter Gill of the Home Office Forensic Science Service, UK, the minimum distance that can be distinguished is on the order of from 1—2% of the band weight.) The match proportion for both bands is the product of the individual match proportions for the two bands, times 2, using the Hardy–Weinberg law: which band is maternal and which is paternal is not known. (Testing a parent of a suspect gives no worthwhile information without also testing a parent of the criminal!)

The Hardy–Weinberg law assumes "random mating" and that the measured band weights are independent. Evett, Werrett, Gill and Buckleton (1989) make it clear that measurement errors are not independent. Many of the factors that contribute to the measurement error for one band also contribute to the measurement error for other bands. So the measurement errors between bands are dependent and, in fact, positively correlated. This dependence is called "band shifting." (Replicates of a single sample carried out and provided to me by Lifecodes indicate a correlation of about 33% for the weights of the two bands on probe D2S44 ($n = 90$) and 70% on probe D17S79 ($n = 120$).)

Laboratories multiply the resulting proportions across all probes used. This assumes that the band weights are independent across probes, which is shown to be questionable by Cohen (1990). It also assumes that measurement errors are independent; this assumption is clearly incorrect because of band shifting. Nevertheless, I will assume independence in this article so I can compare the approach I suggest with the match/binning approach.

Multiplying proportions gives estimates of overall match proportions that tend to be very small. Lander (1989a) cites a case in which the match proportion was calculated to be 1 in 738 trillion. The next section gives an example in which the match proportion is larger than this, but small nonetheless.

3. EXAMPLE: *NEW YORK v. CASTRO*

José Castro, an Hispanic male, was accused in 1987 of murdering Vilma Ponce and her 2-year-old daughter. He was identified as the murderer by Ponce's common-law husband (Lewin, 1989). Lifecodes compared a sample of blood taken from the defendant's watch with that of Vilma Ponce and reported odds of 1:189,200,000. (Lander, 1989a, b, criticized this report on numerous grounds, some statistical.) I will use the band weights reported in this case to show how match proportions are calculated.

Lifecodes used four probes: D2S44, D17S79, DXYS14 and DXY1. The last of these is used to ascertain sex; the results of this particular experiment were questionable (Lander, 1989b). Table 1 gives the band weights (in kb) for the other three probes. The pattern for D17S79 is typical for heterozygotic individuals and single-locus probes: two bands, one inherited from the mother and the other from the father. Regarding D2S44, apparently both Ponce and the individual whose blood was on Castro's watch are homozygotic: the maternal and paternal bands coincide (unless one is missing; see discussion below). Actually, they are what I call "measurably homozygotic": the maternal and paternal fragment lengths are sufficiently close that the laboratory process cannot resolve them as distinct bands.

Lander points out several problems with Lifecodes' use of probe DXYS14. He also argues that "DXYS14 is poorly suited for forensic work, since it can detect anywhere between 2 and 6 bands"; that is, it is a multilocus probe. I think it is "poorly suited" only because good statistical methods are not available for analyzing band weights of multilocus probes; if such methods were available, this probe would be quite powerful. In any case, I will restrict consideration to D2S44 and D17S79.

A reference population is required for computing match proportions. Lifecodes used their Hispanic data base, so I will too, but in Section 7, I will criticize the criteria by which laboratories select populations to be used for reference. Lifecodes' data bases for these two probes are given in Balazs, Baird, Clyne and Meade (1989); their Tables 1(C) and 3(C) give the frequency distributions for "296 Hispanics" for D17S79 and "294 Hispanics" for D2S44. (Actually, these are from about half as many individuals.) These are shown in Figures 1 and 2, in which I count $n = 295$ and $n = 292$, respectively. (I will describe the shaded areas of these histograms presently.) Balazs, Baird, Clyne and Meade gave a nonlinear scale for D2S44; the different widths of the bars in Figure 2 are the result of linearizing their scale.

TABLE 1
Fragment lengths in Castro case

Source/Probe	D2S44	D17S79		DXYS14		
Vilma Ponce	10.162	3.869	3.464	4.855	2.999	1.946
Blood on watch	10.350	3.877	3.541	4.858	2.995	1.957

Consider D2S44. I indicated above that Lifecodes estimates the standard deviation between two measurements to be 0.6% of their average: $(0.006)(10.256) = 0.0615$ (kb). Three s.d.'s is 0.185 kb. The difference between Ponce's band weight and that of the blood on Castro's watch was $10.350 - 10.162 = 0.188$ (kb), which is greater than 0.185, and so this is not a match when assuming $k = 3$. Lifecodes claimed a match.

I will indicate the match proportions reported by Lifecodes, but I will first calculate these proportions using Lifecodes current procedure, which uses the data base proportion within $k = 3$ s.d.'s, or 1.8%, of the victim's band weight: $(0.180)(10.162) = 0.183$ (kb). A match would be declared (using this definition) for any band weight within 10.162 ± 0.183, that is, from 9.979 to 10.345. The

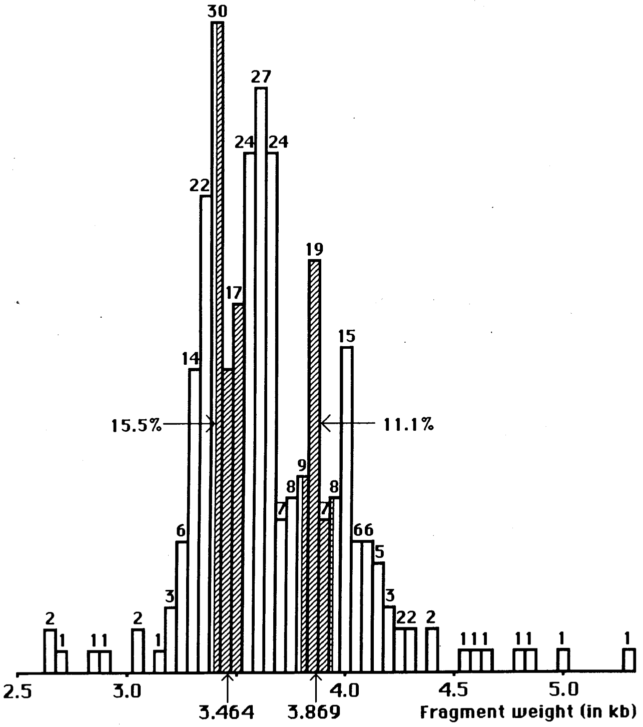

FIG. 1. *Frequency distribution of band weights for probe D17S79, 295 Hispanics. From Figure* 1(*C*) *of Balazs, Baird, Clyne and Mead* (1989).

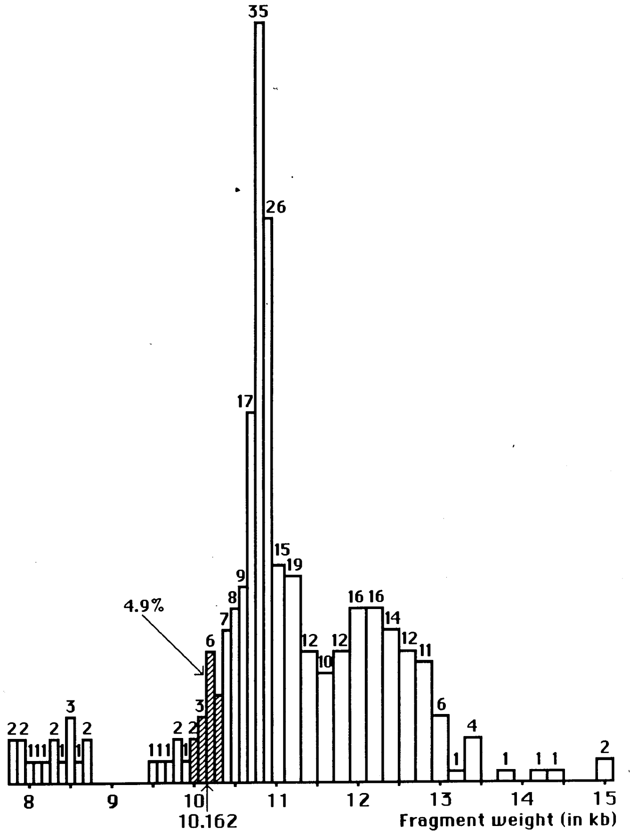

FIG. 2. *Frequency distribution of band weights for probe D2S44, 292 Hispanics. From Figure* 3(*C*) *of Balazs, Baird, Clyne and Mead* (1989); *adapted by making the scale linear and changing heights of bars so frequency is now proportional to area.*

frequency in this interval is the area shaded in Figure 2. Allocating the frequencies in the two outermost categories proportionately gives a total relative frequency of 4.9%. (It is evident that this is an underestimate because their de facto match criterion also included the band weight of the watch sample.) The procedure Lifecodes used at the time of the Castro case (Morris, Sanda and Glassberg, 1989) gave 2.1%. (Using $k = 2/3$ gives 1.4%; since this is not too different from 2.1%, this lends empirical support to Lander's theoretical calculation showing that this procedure is roughly equivalent to using a bin size of $\pm 2/3$ s.d.'s.)

Since they observed only one band on this probe, Lifecodes concluded that both individuals were (measurably) homozygous; under this assumption the overall probe match proportion is the square of the band frequency: $(0.049)^2 \approx 1/420$, or $(0.021)^2 \approx 1/2270$. Lander (1989b) claims that the watch sample may have contained a second, larger band that was not visible on the autoradiogram because the bloodstain sample was small and it may have degraded. He suggested that Lifecodes should have used a *monomorphic* probe (one for which everyone has the same band weight, say one at 15 kb) to see whether their process was able to detect large band weights in the watch sample. Under the circumstances, Lander claims that a more appropriate calculation would be to take twice the match proportion: $2(0.049) \approx 1/10$, or $2(0.021) \approx 1/24$.

I will consider both homozygote and heterzygote frequencies, say p^2 and $2p$. However, while I agree with Lander that the former is too small, his is too large—perhaps much too large. Lander says "the exact formula is $2p - p^2$." I would like to modify his formula to $2p(p + W) - p^2$, where W is the probability that the watch sample's second band is larger than the measured band at 10.350 kb and was not observed because of degradation. This specializes to the two extremes by taking W equal to 0 and $1 - p$. Neither of these extremes is tenable: W is obviously less than $1 - p$ even if 10.350 were the largest band weight observable, which seems unlikely. Referring to Figure 2, the total frequency of band weights near 10.350 and larger shows that $p + W$ is no larger than 90%. Better estimates of W are possible and could be provided by technicians based on their experience. But in preparing for eventualities such as existed in this case, laboratories should estimate W for each possible observed band weight and for various possible qualities and quantities of sample. One way to do this is to conduct experiments designed to estimate the probability of missing a band as a function of its weight (for samples of different quality and quantity) and then averaging this with respect to the frequency distribution of band weights in the population.

Now consider probe D17S79. The average of the larger of the two band weights is 3.873 kb. Three s.d.'s is $3(0.006)(3.873) = 0.070$ (kb). Since this is greater than the difference in band weights, 0.008, these bands match. The match proportion is the shaded area within 3.869 ± 0.070 shown in Figure 1: 11.1%. The procedure Lifecodes used at the time of the Castro case gave 2.3%.

The distance between the smaller band weights on D17S79 is 0.077 kb. This is 3.66 s.d.'s and so this band provides an exclusion for any $k < 3.66$. Still, Lifecodes called a match and calculated a match proportion of 6.8% for this band. Using $k = 3$ gives the shaded area in Figure 1: 15.5%.

Combining these proportions as described above gives $2(0.023)(0.068) \approx 1l04/320$, *and* $2(0.111)(0.155) \approx 1/29$. Assume independence of probes D2S44 and D17S79. Using the Lifecodes match proportions, the estimated frequency of the observed patterns on these probes in the Hispanic population is $(1/2270)(1/320) \approx 1/725{,}000$ (homozygosity in D2S44) and $(1/24)(1/320) \approx 1/7610$ (heterozygosity and missing band in D2S44). The former was claimed by Lifecodes (it is opposed to the figure 1/189,200,000 mentioned earlier, which was based on all four probes). Using $k = 3$, the estimated frequency of this pattern on these probes in the Hispanic population is $(1/420)(1/29) \approx 1/12{,}200$ (homozygosity in D2S44) and $(1/10)(1/29) \approx 1/290$ (heterozygosity and missing band in D2S44).

A match proportion is an indirect measure of the suspect's guilt because its calculation *assumes the suspect is innocent*. The approach of the next section addresses a direct calculation of the probability that the suspect is guilty. I will apply this calculation to the Castro case in Section 7.

4. LIKELIHOOD RATIO OF GUILT TO INNOCENCE: SINGLE-BAND CASE

The method described in this section was first applied to a problem in forensic identification (that of glass) by Lindley (1977).

Consider a single DNA band weight. This is unrealistic (except possibly in paternity cases; see Section 8) because even single locus probes have pairs of bands, one maternal and the other paternal. But the methodology carries over to pairs of bands and also to multiple single-locus probes. Section 5 describes this extension for the independent case.

Suppose β is the true weight of the crime sample band and α is the true weight of the suspect's band. In the scenario considered in the previous section, if the suspect is guilty then $\beta = \alpha$. If the suspect is innocent then β may or may not equal α. They will be equal if the suspect's identical twin is guilty, and possibility in other cases as well. But for hypervariable loci, they are usually different when the suspect is innocent. The values of β and α are not observable since measuring band weights is subject to error. Let y and x be the measured band weights of the crime sample and the suspect, respectively. I will discuss the distribution of measurement error below.

The question of interest is whether the suspect is guilty (G); that is, whether the suspect's and crime sample's DNA fragments are from the same person. A way to address this question is to find the probability of G conditioned on the available evidence. This requires Bayes' theorem. It is convenient to partition the evidence into two pieces, $\boldsymbol{X}$ and E. Here, $\boldsymbol{X}$ is the set of measured band weights of suspect, crime sample and n other individuals, $\boldsymbol{X} = (x, y, z_1, z_2, \ldots, z_n)$. And E is all other evidence, such as eyewitness accounts. Letting I (for innocent) stand for the complement of G, Bayes' theorem says that the posterior odds of guilt are proportional to the prior odds of guilt:

$$(1) \qquad \frac{P(G \mid \boldsymbol{X}, E)}{P(I \mid \boldsymbol{X}, E)} = R\frac{P(G \mid E)}{P(I \mid E)},$$

where R is the Bayes factor (or likelihood ratio):

$$R = \frac{P(\boldsymbol{X} \mid G, E)}{P(\boldsymbol{X} \mid I, E)}.$$

I will take background evidence E as understood throughout and suppress it from the notation. With this convention, the likelihood ratio is

$$R = \frac{P(\boldsymbol{X} \mid G)}{P(\boldsymbol{X} \mid I)}.$$

I will focus on the likelihood ratio in this section and consider the prior odds in Section 6.

I will temporarily consider a rather artificial setting. Prior information indicates that if the suspect is innocent then one of the other n individuals is guilty, each with probability $1/n$. This means that β equals at least one of $\alpha, \alpha_1, \alpha_2, \ldots, \alpha_n$, the true band weights of the suspect and of the other n members of the population. This assumption would be appropriate in a closed community of $n + 1$ individuals, all of whom are tested. (It is also appropriate when the sample size n is sufficiently large that there is effectively no sampling variability. But it is not usually realistic. My main purpose in considering it is that the calculations are easy and instructive.)

I will assume that the measurement errors are independent. Baird et al. (1986) suggest that they are normally distributed with s.d. proportional to the mean. Independent replicates supplied to me by Baird support the assumption of normality. They are equally supportive of the assumption of lognormally distributed errors with constant s.d.; because the s.d. is small, these two assumptions are very similar. For mathematical convenience I will assume that errors are lognormally distributed:

$$\log y \sim \boldsymbol{N}(v, c^2), \quad \log x \sim \boldsymbol{N}(\mu, c^2),$$
$$\log z_i \sim \boldsymbol{N}(\mu_i, c^2).$$

(Until now I have made no assumptions about the relationship between observed band weights such as y and actual band weights such as β. But if, for example, β is assumed to be the mean of y, then $v = \log \beta - \mathrm{c}^2/2$.)

Let m and s be the average and s.d. of $\log x$ and $\log y$:

$$m = \frac{\log(xy)}{2} \quad \text{and} \quad s = \frac{|\log(x/y)|}{2}.$$

Also, let s_i be the s.d. of $\log z_i$ and $\log y$:

$$s_i = \frac{|\log(z_i/y)|}{2}.$$

Since G means that $\mu = v$, the likelihood ratio is

$$R = \frac{p(\boldsymbol{X} \mid \mu = v)}{p(\boldsymbol{X} \mid I)}, \tag{2}$$

where I write lower case p to indicate that these are densities. To give an example with rather easy calculations I will make the usually unrealistic assumption that the μ's are independent with improper uniform prior distribution on $(-\infty, +\infty)$. In this case the numerator of (2) is

$$\begin{aligned} & p(\boldsymbol{X} \mid \mu = v) \\ &= \int p(\boldsymbol{X} \mid \mu = v, \mu, \mu_1, \ldots, \mu_n) \\ &\qquad \cdot dF(\mu, \mu_1, \ldots, \mu_n) \\ &= \int \frac{1}{2\pi c^2 xy} \exp\left\{ -\frac{1}{2c^2} \left[(\log x - \mu)^2 \right.\right. \\ &\qquad\qquad \left.\left. + (\log y - \mu)^2 \right] \right\} d\mu \\ &\qquad \cdot \prod_{i=1}^{n} \int \frac{1}{c z_i \sqrt{2\pi}} \\ &\qquad\qquad \cdot \exp\left\{ -\frac{1}{2c^2} (\log z_i - \mu_i)^2 \right\} d\mu_i \\ &= \frac{1}{2\pi c^2 xy} \int \exp\left\{ -\frac{1}{c^2} \left[s^2 + (m - \mu)^2 \right] \right\} d\mu \\ &\qquad \cdot \prod_{i=1}^{n} \left(\frac{1}{z_i} \right) \\ &= \frac{1}{cxy\sqrt{\pi}} \exp\{ -(s/c)^2 \} \prod_{i=1}^{n} \left(\frac{1}{z_i} \right) \\ &= K \exp\{ -(s/c)^2 \}. \end{aligned} \tag{3}$$

The denominator of (2) has a similar form:

$$\begin{aligned} & p(\boldsymbol{X} \mid I) \\ &= \sum_{i=1}^{n} p(\boldsymbol{X} \mid \text{person } i \text{ guilty}) \\ &\qquad \cdot P(\text{person } i \text{ guilty} \mid I) \\ &= \sum_{i=1}^{n} p(\boldsymbol{X} \mid \mu_i = v) \frac{1}{n} \\ &= \frac{K}{n} \sum_{i=1}^{n} \exp\{ -(s_i/c)^2 \}. \end{aligned} \tag{4}$$

So (1) becomes

$$R = \frac{\exp\{ -(s/c)^2 \}}{n^{-1} \sum_{i=1}^{n} \exp\{ -(s_i/c)^2 \}}. \tag{5}$$

Equation (5) is especially simple (and in many cases it gives answers similar to the usually more realistic (9) that will be developed below). It shows clearly the effect of single-band DNA evidence. The maximum value of R occurs when the evidence is most incriminating: $x = y$. If x and y are far apart then R is small, decreasing exponentially as the square of the log of their ratio. For example, if the standard deviation s of $\log x$ and $\log y$, which is half the log of their ratio, is twice that of the measurement standard deviation c, then R is only $\exp\{-4\}$ (or about 1.8%) of its maximum. And if $s = 3c$, so the crime sample is rather inconsistent with the suspect, R is only 0.012% of its maximum.

If x and all the z's are equidistant from y, then $s = s_1 = s_2 = \ldots = s_n$; in this case (5) gives $R = 1$ and there is no evidentiary value in data $\boldsymbol{X}$. If x and z_1, say, are equidistant from y, and other z's are much further away, then $s = s_1$ is much smaller than the other s_i's and so $R \approx n$. And R is effectively 0 if x and y are quite discrepant in comparison with some of the z_i's. But while R can be small, it cannot be zero. So if other probes are consistent with guilt, the suspect may still have a large overall likelihood ratio (see Section 5) and a correspondingly large probability of guilt. Section 7 illustrates (5) using the Castro example.

The exponential nature of (5) stems from the assumption of normal measurement errors. If the actual distribution of errors has larger tails than does the normal, then the analogous likelihood ratio will be less affected by a large difference between x and y. In practice, laboratories should do a careful analysis of duplicate measurements to estimate the underlying error distribution. If its tails are as small as or smaller than the normal, then the normality assumption will give reasonably accurate conclusions. But the normal will give poor results if there are outliers. (On the other hand, the normal is much more robust than a ±k s.d. criterion.)

It is usually unrealistic to assume that the entire set of possible criminals has been tested. More typically, the tested individuals represent a sample from some population; perhaps n is a few hundred, not large enough to be confident that the sample frequencies adequately reflect the population frequencies. Take an extreme example. Suppose y and x are near each other, but in a region containing *none* of the z's. Then R will be extremely large, say 10^{10}. If we are certain that the criminal is among the $n + 1$ individuals tested and that our assumed error distribution is correct (I'm speaking hypothetically: We're never *certain*!), then this enormous value of R is reasonable and we're confident we've solved the crime.

In the example of the previous paragraph, suppose the n individuals are a *sample* from a population containing the criminal. The region containing y and x is quite clearly a rather sparse region in the population as well, but because of sampling variability there may be population members in the region who have not been tested. While it may be reasonable to say that the likelihood of the suspect's guilt is 10 billion times that of the sample members combined, it is preposterous to make such a conclusive statement when extrapolating to the larger population. For suppose the next person sampled happens to be within a standard deviation or so of y. Then R will decrease by about eight orders of magnitude. Such an observation deserves to be influential, but not that influential!

In line with the previous paragraph, I will drop the assumption that $(\mu, \mu_1, \dots, \mu_n)$ is the population of band weights. I will now assume that each μ is a band weight selected from the appropriate population and that $(\mu, \mu_1, \dots, \mu_n)$ is a representative sample from this population. For example, if a murder takes place in an isolated, highly inbred town, then the sample is supposed to be representative of the town; in particular, it would then be unreasonable to use a sample from a larger population of which the town is a subset. Making the realistic assumption that G, the suspect is guilty, and the reference data set $\boldsymbol{Z} = (z_1, \dots, z_n)$ are independent, likelihood ratio (2) can be written instead as

$$R = \frac{p(x, y \mid \mu = v, \boldsymbol{Z})}{p(x, y \mid \boldsymbol{Z})}. \tag{6}$$

In analogy with (3) and (4),

$$\begin{aligned} p(x, y \mid \mu = v, \boldsymbol{Z}) &= \int p(x, y \mid \mu, v, \mu = v)\, dH(\mu \mid \boldsymbol{Z}) \\ &= \int \frac{1}{2\pi c^2 xy} \exp\left\{ -\frac{1}{2c^2}\left[(\log x - \mu)^2 \right.\right. \\ &\qquad \left.\left. + (\log y - \mu)^2\right]\right\} dH(\mu \mid \boldsymbol{Z}), \end{aligned} \tag{7}$$

and

$$\begin{aligned} p(x, y \mid \boldsymbol{Z}) &= \int \frac{1}{cx\sqrt{2\pi}} \\ &\quad \cdot \exp\left\{ -\frac{1}{2c^2}(\log x - \mu)^2 \right\} dH(\mu \mid \boldsymbol{Z}) \\ &\quad \cdot \int \frac{1}{cy\sqrt{2\pi}} \\ &\quad \cdot \exp\left\{ -\frac{1}{2c^2}(\log y - v)^2 \right\} dH(v \mid \boldsymbol{Z}), \end{aligned} \tag{8}$$

where $H(\cdot \mid \boldsymbol{Z})$ is the conditional distribution of the population of actual band weights given the sample of observed band weights. The problem is to find $H(\cdot \mid \boldsymbol{Z})$, or a suitable estimate of it.

It would not be appropriate to use the empirical distribution of the logs of the z's to estimate $H(\cdot \mid \boldsymbol{Z})$ in (7) and (8). First, the observed values of $\log z_i$ do not consider measurement error. Second, such an estimate does not consider that Z is a sample and not the population. Finding (7) or (8) can be accomplished using hierarchical Bayesian methods (Berger, 1986). In particular, Ferguson (1983), Lo (1984) and Kuo (1986) give approaches to density estimation using Dirichlet process priors.

I will use a simple-minded density estimation approach. A way to account for measurement error is to use normal kernels and estimate $H(\cdot \mid \boldsymbol{Z})$ as

$$H_1(\cdot \mid \boldsymbol{Z}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{N}(\log z_i, c^2).$$

There are several ways to account for sampling variability. Of special concern is the possibility that the sample values underrepresent the population proportion near x. Such underrepresentation inflates R and so can greatly exaggerate the evidence in favor of G when x is near y. This problem is especially severe when *none* of the z's are near x. One possible remedy is to include a uniform base to the density estimate, as follows:

$$\begin{aligned} H_2(\cdot \mid \boldsymbol{Z}) &= \frac{n^*}{n + n^*} \boldsymbol{U}(N_1, N_2) \\ &\quad + \frac{1}{n + n^*}\sum_{i=1}^{n} \boldsymbol{N}(\log z_i, c^2), \end{aligned}$$

where $\boldsymbol{U}(N_1, N_2)$ is a uniform distribution on N_1 and N_2, a range chosen sufficiently large to include the logs of all possible band weights. The parameter n^* is the number of observations to be spread out uniformly from N_1 to N_2. This particular combination has the characteristic that it tends to $H_1(\cdot \mid \boldsymbol{Z})$ for large n, which is appropriate.

Still another way of accounting for sampling variability is to smooth each data point more than necessary for $H_1(\cdot \mid \boldsymbol{Z})$. For example, multiplying the standard deviation by b at each data point gives

$$H_3(\cdot \mid \boldsymbol{Z}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{N}(\log z_i, (bc)^2).$$

As I indicated in discussing H_1, smoothing parameter $b = 1$ accounts for measurement error in the data base band weights. Since measurement error is always present, b should be at least 1. Generally, it should be greater than 1 and it should be larger for smaller n. If n is so large that the sample frequencies are effectively the same as the population frequencies, then b could be set to 1.

More generally than all the above:

$$H_4(\cdot \mid \boldsymbol{Z}) = \frac{n^*}{n+n^*} U(N_1, N_2) + \frac{1}{n+n^*} \sum_{i=1}^{n} \boldsymbol{N}(\log z_i, (bc)^2).$$

I will use $H_3(\cdot \mid \boldsymbol{Z})$—which is the same as $H_4(\cdot \mid \boldsymbol{Z})$ with $n^* = 0$—although extensions to general n^* are straightforward.

After much algebra in evaluating (7) and (8) with $H_3(\cdot \mid \boldsymbol{Z})$ in place of $\mathrm{H}(\cdot \mid \boldsymbol{Z})$, likelihood ratio (6) becomes

$$(9) \qquad R = \frac{Q_2(m) \exp\{-(s/c)^2\}}{Q_1(\log x) Q_1(\log y)},$$

where, for $j = 1, 2$,

$$Q_j(u) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\sqrt{1+jb^2}} \cdot \exp\left\{ -\frac{j}{2c^2} \frac{(u - \log z_i)^2}{1+jb^2} \right\}.$$

(These sums are trivial when calculating with a computer.) It is not as easy to see how (9) works as it is to understand (5); at least in some cases, such as the Castro example (see Section 7), the two give quite similar results.

A consideration in calculating (9) is choosing the smoothing parameter b. When n is small then b should be large, say around 5, and b should tend to 1 as $n \to \infty$. If x and y are close to each other in a region in which there are few of the z_i's, then the choice of b is crucial (see the example below), with R being larger for b nearer 1. But if x is in a region with at least moderate frequency, then the choice of b is not as important (as in each of the three bands considered in the Castro example of Section 7).

To illustrate (9), consider the frequency distribution of Hispanics for probes D17S79 and D2S44 given in Figures 1 and 2. Figures 3 and 4 show the corresponding smoothed version of these distributions; each figure shows two smoothings, one for $bc = 0.006$ and the other for $bc = 0.03$.

Figure 5 is a contour plot of R as given in (9); the axes are the average band weight $(x + y)/2$ and difference in band weights $(x - y)/2$; $c = 0.006$ and $b = 5$. This plot gives approximate values of R; for example, the two x's in Figure 5 are the data points in the Castro example of Section 7 and correspond to $R = 0.343$ and $R = 16.1$ (cf. Table 3).

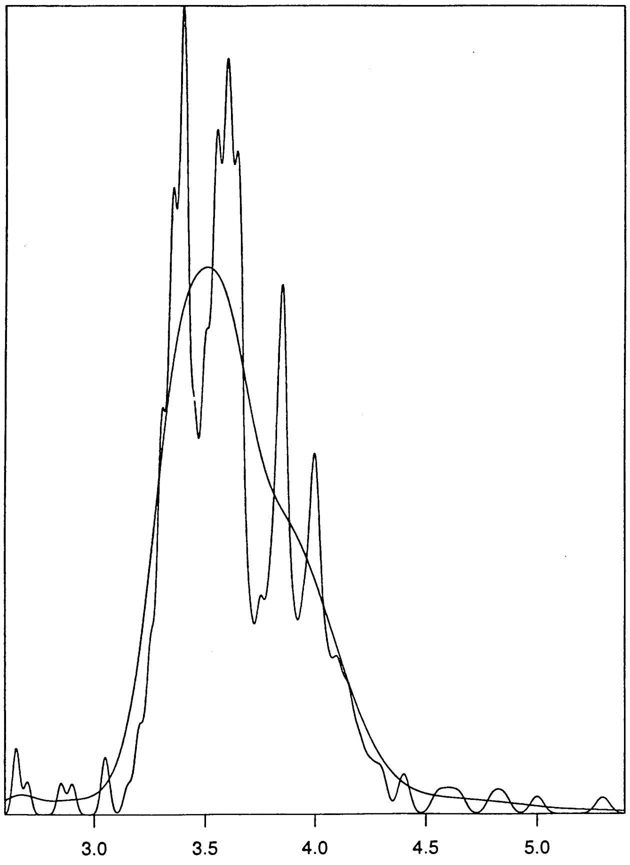

FIG. 3. *Smoothed versions of histrogram in Figure* 1; $b = 1$ *and* $b = 5$, *with* $c = 0.006$ *in both.*

The larger the value of R, the stronger the evidence of guilt; contour $R = 1$ in Figure 5 indicates the locus of points where evidence $\boldsymbol{X}$ is neutral, favoring neither guilt nor innocence. When x is in a region of high frequency (cf. Figure 1 or Figure 3), the maximum value of R is smaller: inferences are not as conclusive when the observed weights are common in the population. And R is larger when x is in a region of low frequency. Also, of course, R decreases as s increases (for fixed m).

To see the effect of b when both x and y are in a region of low frequency, consider probe D2S44 (Figure 4). Suppose $x = y = 9.1$ kb. This is pretty convincing evidence for two reasons: (1) $x = y$ and (2) there are no weights in the reference sample that are nearby. Assuming $c = 0.006$ and $b = 1$ (the wiggly curve in Figure 4), the value of R in (9) is 961,000. On the other hand, the smoothing that results from assuming $b = 5$ (the smoother curve in Figure 4) partially fills in the hole near 9.1 kb and R is substantially decreased to 506. The latter is much more reasonable since, for one thing, the sample size is less than 300.

A way to smooth where it matters most is to include the band weights of the crime sample (or

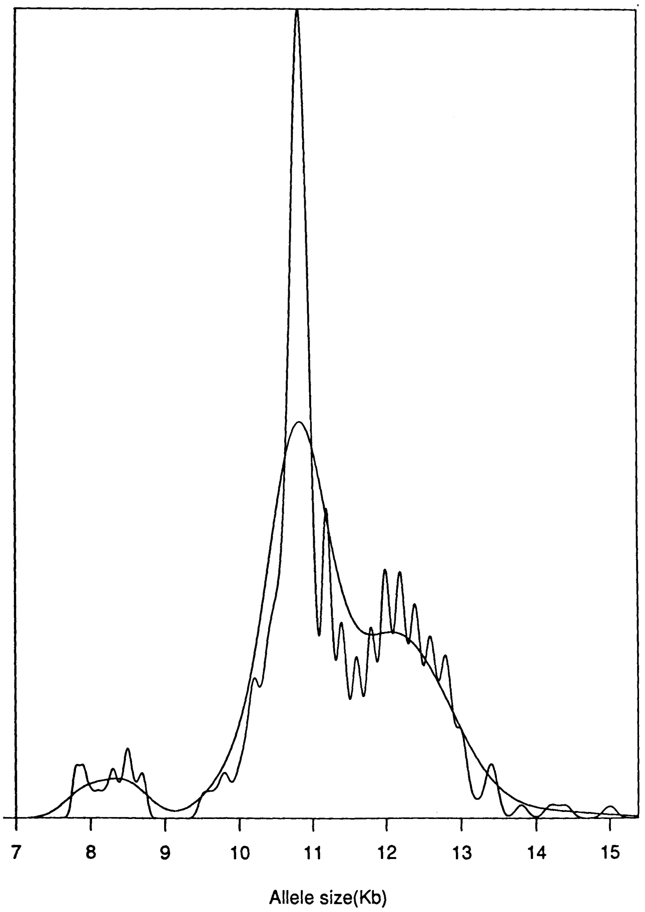

FIG. 4. *Smoothed versions of histrogram in Figure* 2; $b = 1$ *and* $b = 5$, *with* $c = 0.006$ *in both.*

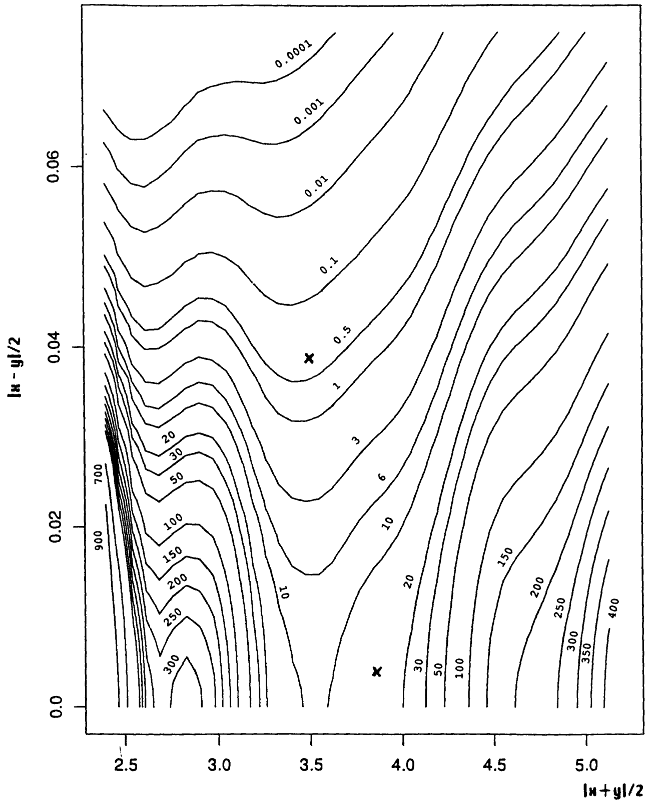

FIG. 5. *Contour plot of likelihood ratio* (9) *in band weight rather than log band weight; probe* = *D17S79, reference data set* = *Hispanics from Figure* 1, $c = 0.006$, $b = 5$. (*Two x's described in Section* 7.)

suspect sample, but not both) in the reference data base. This has the effect of making R smaller, and therefore it is more conservative in the sense of being weaker evidence against the suspect. In the $x = y = 9.1$ kb example for D2S44, adding band weight 9.1 kb to the data base gives $R = 338$ for $b = 1$ and $R = 344$ for $b = 5$. I recommend this procedure for standard practice, but it makes little difference in the Castro example and so I have not used it.

5. INDEPENDENT BANDS AND INDEPENDENT PROBES

I indicated in the previous section that the single-band case is not usually applicable in criminal cases because for any given single-locus probe, each individual has two distinct bands, one maternal and the other paternal (except for homozygotes, in which the two bands coincide). However, the calculations extend to independent bands of a single-locus probe and to multiple independent single-locus probes in a straightforward way. (As I indicated in Section 2, measured band weights may be highly correlated. Evett and Pinchin of the Home Office Forensic Science Service, UK, and I are extending the likelihood ratio approach to handle the dependent case.)

Consider two bands of a single-locus probe. Suppose x_1 and x_2 are the measured weights of the crime sample DNA fragments, and y_1 and y_2 are the measured weights of the suspect's fragments. If the suspect is guilty then there are two possibilities: either the band with measured weight x_1 is paired with y_1 or with y_2. Let R_{jk} be the likelihood ratio for the pair (x_j, y_k), calculated as in the previous section at (5) or (9). Since the two possible pairings are equally likely, the likelihood ratio for the probe is

$$(10) \qquad R = \frac{1}{2}R_{11}R_{22} + \frac{1}{2}R_{12}R_{21}.$$

When the DNA evidence consists of both bands of a single-locus probe, Bayes' theorem can be used as the previous section, with R given by (10).

Two special cases are of interest in calculating (10). When x_1 is much larger than x_2 (in terms of measurement s.d.'s) and y_1 is much larger than

y_2, then R_{12} and R_{21} will be small; so in this case

$$R \approx \frac{1}{2} R_{11} R_{22}.$$

Also, when $x_1 = x_2$ and $y_1 = y_2$ (both crime sample and suspect are measurably homozygotic), then all four single-band likelihood ratios are equal; in this case

$$R \approx R_{11}^2.$$

The extension to several independent single-locus probes is even easier: The likelihood ratios for the individual probes are calculated as just indicated above and multiplied together.

6. PROBABILITY OF GUILT

Calculations in Section 4 and 5 deal with the likelihood ratio of guilt to innocence. It is appropriate to present this likelihood ratio and describe its calculation to a court. But its inverse is not the probability of innocence, nor is there a transformation of it alone that gives the probability of guilt or innocence. To use a likelihood ratio, the jury must understand how it combines with other evidence in the case.

Calculating the probability of guilt in view of the DNA data and other evidence E requires Bayes' theorem (1). The "other evidence" enters through the prior probability $P(G)$ or the prior odds, $P(G)/P(I)$. I have suppressed E in writing $P(G)$, e.g., for notational simplicity, but it is logically correct to evaluate all probabilities discussed in this paper as conditional on E. A job of the forensic statistician is to explain to a court how to convert $P(G \mid E)$ into $P(G \mid \boldsymbol{X}, E)$. This is easier to accomplish using odds: Multiply $P(G \mid E)/P(I \mid E)$ by R to obtain $P(G \mid \boldsymbol{X}, E)/P(I \mid \boldsymbol{X}, E)$.

A more challenging problem is to help jurors assess $P(G \mid E)$. Comparing preferences for the prospect G as compared with well-understood bets (coins and dice) may help in this endeaver (see DeGroot, 1970, Chapter 6). At oft-made suggestion is that $P(G \mid E)$ should be one over the size of some population. This may be reasonable if a population of possible criminals can be identified, and if the suspect is exchangeable with other members of this population assuming E. (If the suspect is exchangeable with them, they must all be suspects.) But to take the population to be that of California, say, just because the crime took place in California and the suspect lived there, is much too conservative: it replaces E (which includes information concerning why the suspect was tested in the first place) with the much weaker evidence that the suspect is a Californian.

In using Bayes' theorem in cases of disputed parentage, blood banks set the prior odds of paternity equal to 1. This assumes that the alleged father has probability 1/2 of being the true father apart from the genetic evidence. Such an assumption is objectionable in paternity cases (Berry and Geisser, 1986), but arbitrarily selecting 1/2 or any other particular probability of guilt in criminal cases would be wholly inappropiate—perhaps even criminal! It would misrepresent the evidence and subsume the roles of judge and jury. Even though such an assumption is not made explicitly in criminal cases, it is difficult to state a likelihood ratio without it being understood to mean posterior odds.

An important issue in assessing $P(G \mid \boldsymbol{X}, E)$ is that DNA evidence $\boldsymbol{X}$ and other evidence E can never really be independent for a juror, and their dependence is difficult to evaluate. Part of the background information always present is that the defendant has been charged with a crime and that the prosecution had what they perceive to be a case worth prosecuting. The prosecution would probably have dropped the case if the DNA evidence had been exclusionary. So $\boldsymbol{X}$ and E are intertwined. It is important for the prosecution to make it clear to the jury why the case came to trial. For example, if the police had planned to test everyone in a town until finding a DNA match and the suspect happened to be first one tested, then the probability of guilt separate from the DNA evidence is *at most* one over the population of the town. At the other extreme would be a suspect who was tested because several eyewitnesses put him (and no others) at the scene of the crime. I don't mean that the prior probability of guilt in the latter case is necessarily near 1, but that other considerations being the same it should be substantially greater than the prior probability of guilt in the former case.

Match/binning does not lend itself well to using Bayes' theorem, but I want to compare it with the likelihood ratio technique developed in Sections 4 and 5. So I will suggest how a match proportion might be interpreted as a likelihood ratio. As in Section 4,

$$\frac{P(G \mid \boldsymbol{X}, E)}{P(I \mid \boldsymbol{X}, E)} = \frac{P(\boldsymbol{X} \mid G)}{P(\boldsymbol{X} \mid I)} \frac{P(G \mid E)}{P(I \mid E)}.$$

Now, instead of $\boldsymbol{X} = (x, y, z_1, z_2, \ldots, z_n)$, take $\boldsymbol{X}$ to be either $\boldsymbol{X}_M = (\text{match}, y, z_1, z_2, \ldots, z_n)$ or $\boldsymbol{X}_N = (\text{not match}, y, z_1, z_2, \ldots, z_n)$. Take the likelihood $P(\boldsymbol{X}_N \mid G)$ to be 0. Take the likelihood $P(\boldsymbol{X}_M \mid G)$ to be 1 and $P(\boldsymbol{X}_M \mid I)$ to be the proportion of z's close to y: the match proportion. Then the "likelihood ratio" is either

$$\frac{P(X_N \mid G)}{P(X_N \mid I)} = 0, \quad \text{or}$$

$$\frac{P(X_M \mid G)}{P(X_M \mid I)} = \frac{1}{\text{match proportion}}.$$

In the next section I will use the latter expression to compare match/binning with the likelihood ratio procedure developed in Section 4 and 5 in the context of the Castro example.

7. EXAMPLE: *NEW YORK v. CASTRO* (REVISITED)

As described in Section 3, suspect Castro had a bloodstain on his watch that may have been the blood of the victim Ponce. I assumed in Section 4 that measured weights of DNA fragments of the suspect and of the crime sample were available and we wanted to know whether they were from the same person. In this example, the suspect may be innocent even if the blood on his watch is that of the victim. On the other hand, he may be guilty even if the blood on his watch is not that of the victim. Handling these possibilities is easy, but I will avoid the extra algebra by assuming $P(\text{blood on watch is Ponce's} \mid G) = 1$ and $P(\text{blood on watch is Ponce's} \mid I) = 0$. This lets me illustrate the calculations described in the previous sections letting x refer to the blood on the watch and y refer to that of the victim, Vilma Ponce (vice versa gives the same answer).

We need a sample $Z = (z_1, z_2, \ldots, z_n)$. It is wrong to assume the race of the suspect as the reference population. The appropriate population is that of the blood on the watch, assuming it is *not* Ponce's. In particular, whether Ponce and Castro are Hispanic is irrelevant (although the jury may feel it likely that blood on the watch of an Hispanic is that of an Hispanic). (The blood on his watch was apparently found not to be Castro's own. Indeed, because of the problems pointed out in Lander, 1989a, b, this was the only DNA evidence that the judge allowed to be introduced in the trial.) Evidence that suggests a probability distribution concerning the race of a person who might have spilled blood on Castro's watch is relevant; in particular, that distribution can be used for averaging the various likelihood ratios (or match proportion when using match/binning). In the absence of such information, or perhaps in any case, calculations should be made for each possible reference population. Since Lifecodes used their Hispanic data base as the reference sample, I will too.

Table 2 shows the likelihood ratio R calculated from (5) for each of the three bands from D2S44 and D17S79 given in Table 1; the reference samples Z are given in Figures 1 and 2. I show this table to illustrate (5), but I remind you that it applies when there is a fixed population containing the criminal and the entire population is $(x, z_1, z_2, \ldots, z_n)$. Table 2 considers various values of c from 0.0042 to 0.03. As described at the end of the previous section, entries should be compared with Lifecodes' calculations of $1/0.021 \approx 48$, $1/0.023 \approx 43$, and $1/0.0068 \approx 150$. Consider the first entry in Table 2: $R = 0.289$. There are two reasons it is so small. The more important is that the observed s.d. of log x and log y ($s = 0.0092$) is much larger than the assumed laboratory s.d. ($c = 0.0042$). The other reason is that there are weights in the data set Z that are much closer to y than is x. The last two rows of Table 2 show the appropriate product of the likelihood ratios for the three bands; the penultimate row assumes homozygosity and so the entries are comparable with 725,000 using Lifecodes' method, while the last row assumes heterozygosity and so the entries are comparable with 7610 using Lifecodes' method. Assuming

TABLE 2

Likelihood ratios for New York v. Castro calculated using (5) (x = *Ponce's band weight*; y = *band weight of watch sample*)

Probe/Band		c							
		0.0042	0.006	0.008	0.010	0.015	0.020	0.025	0.030
D2S44	$x = 10.162$ $y = 10.350$	0.0289	2.20	4.43	5.46	5.02	3.98	3.29	2.84
D17S79	$x = 3.869$ $y = 3.877$	20.0	14.1	11.1	9.19	6.47	4.97	3.98	3.30
D17S79	$x = 3.464$ $y = 3.541$	0.013	0.304	1.01	1.60	2.15	2.11	1.96	1.82
Combination: D2S44 homozygous		0.010	10	110	220	180	83	42	24
Combination: D2S44 heterozygous		0.018	2.4	12	20	17	10	6.4	4.3

Lifecodes' estimate of $c = 0.0042$ (the leftmost column in Table 2), the entries tend to be much smaller than Lifecodes' figures. In fact, assuming homozygosity in D2S44, the reference sample has about $1/0.010 \approx 100$ times *more* likelihood than does Ponce of accounting for the blood on Castro's watch!

Table 3 is similar to Table 2, except that in Table 3 the likelihood ratio for each band is calculated from (9). The table shows the effect of the smoothing parameter b in relation to c. Calculations for $b < 1$ are shown for illustration only; in practice b should be at least 1. All three bands happen to be in regions of moderately high density. So the s.d. of the smoothing kernel, bc, has little effect on the overall likelihood ratio. As indicated in Section 4, had one or more bands occurred in sparse regions then smoothing would have much more effect, with a greater amount of smoothing making large R's smaller. Also as indicated in Section 4, the two x's on Figure 5 show the likelihood ratios (16.1 and 0.343) for the two bands of probe D17S79 and assuming $c = 0.006$ and $bc = 0.03$.

One conclusion from Table 3 is that if the blood on Castro's watch is indeed that of Vilma Ponce (Castro has since pleaded guilty!) then Lifecodes has underestimated their laboratory measurement standard deviation. For example, assuming homozygosity of D2S44 and using $c = 0.0042$ (Lifecodes's estimate) and $bc = 0.03$, the data are about $1/0.02 \approx 50$ times more likely when assuming *innocence* than when assuming guilt! Using $c = 0.006$, which is my estimate from the data Lifecodes used to derive $c = 0.0042$, increases the overall likelihood ratio (to about 16) by three orders of magnitude. But the table suggests that even this estimate may be too small.

8. PROBABILITY OF PATERNITY

The above development adapts easily to calculating the probability that an alleged father of a child is the true father. Now the evidence includes the weights of the child's, mother's and alleged father's DNA segments. (The following discussion applies with easy modifications when the mother's is absent.) The likelihood ratio is called the "paternity index" in cases of disputed paternity (Berry and Geisser, 1986); the event G now means that the alleged father is the true father.

Consider a single probe. Gjertson, Mickey, Hopfield, Takenouchi and Terasaki (1988) consider all possibilities. I will consider only the simple case in which mother, child and alleged father have two widely separated bands and that one of the child's

TABLE 3

Likelihood ratios for New York v. Castro calculated using (9) (*x = Ponce's band weight*; *y = band weight of watch sample*)

		c							
Probe/Band	bc	0.0042	0.006	0.008	0.010	0.015	0.020	0.025	0.030
	0.0042	0.334	2.51	5.34	6.91	6.99	5.39	4.04	3.19
D2S44	0.006	0.347	2.56	5.39	6.92	6.91	5.33	4.02	3.19
$x = 10.162$	0.010	0.357	2.60	5.37	6.78	6.62	5.16	3.97	3.20
$y = 10.350$	0.020	0.317	2.28	4.67	5.86	5.82	4.78	3.88	3.25
	0.030	0.272	1.97	4.07	5.17	5.33	4.57	3.85	3.30
	0.0042	17.6	13.3	10.8	9.16	6.54	5.06	4.09	3.37
D17S79	0.006	18.4	13.9	11.0	9.22	6.53	5.05	4.08	3.36
$x = 3.869$	0.010	19.7	14.6	11.4	9.34	6.54	5.04	4.05	3.34
$y = 3.877$	0.020	21.6	15.8	12.0	9.74	6.61	4.98	3.97	3.29
	0.030	22.1	16.1	12.2	9.83	6.58	4.93	3.93	3.26
	0.0042	0.019	0.363	1.10	1.65	2.04	1.97	1.84	1.73
D17S79	0.006	0.018	0.343	1.07	1.63	2.05	1.98	1.85	1.74
$x = 3.464$	0.010	0.016	0.324	1.04	1.62	2.09	2.03	1.89	1.76
$y = 3.541$	0.020	0.016	0.326	1.06	1.68	2.23	2.17	2.00	1.85
	0.30	0.017	0.343	1.12	1.78	2.36	2.30	2.11	1.94
	0.0042	0.019	15	170	360	330	140	61	30
Combination: D2S44	0.006	0.019	16	170	360	320	140	61	30
homozygous	0.010	0.020	16	170	350	300	140	60	30
	0.020	0.017	13	140	280	250	120	60	32
	0.030	0.014	11	110	230	220	120	61	34
	0.0042	0.0028	3.1	16	26	23	13	7.6	4.6
Combination: D2S44	0.006	0.028	3.1	16	26	23	13	7.6	4.7
heterozygous	0.010	0.028	3.1	16	26	23	13	7.6	4.7
	0.020	0.028	2.9	15	24	21	13	7.7	4.9
	0.30	0.025	2.7	14	23	21	13	8.0	5.2

bands matches unambiguously with one and only one of the mother's. The child's other band is paternal. Suppose y is the measured weight of this latter band.

The alleged father has two bands. If he is the actual father, then the child has probability 1/2 of inheriting each. Assuming that the proportion of homozygotes is negligible, this probability is the same for essentially any possible father. So it will cancel in calculating the likelihood ratio of paternity versus nonpaternity. Let x denote the measured weight of the alleged father's band that is closer to y (and suppose that the other is much further away). Then the question is whether v (the true weight of the paternal band) is the same as μ (the true weight of the alleged father's band), just as in Section 4. So the calculation proceeds just as in Section 4. In cases of disputed paternity there is no multiplying of likelihood ratios for two bands on the same probe. So there is one calculation as in (5) or (9) for each single-locus probe.

If several independent single-locus probes are used, the likelihood ratios for each can be multiplied, just as in Section 5. But for the same reasons as in criminal cases, the assumption of independence is suspect.

The appropriate reference population is again problematic. Blood banks and other facilities that do paternity testing use the race of the *alleged* father to determine the reference population. They should use the race of the *true* father, assuming he is not the alleged father. If the latter is unknown, or contested, then calculations should be made for each race and perhaps also averaged in some way.

9. CONCLUSION AND DISCUSSION

The match/binning approach currently used by forensic laboratories is a rather crude but satisfactory procedure for inferring degree of match in a *scientific* setting. The scientist understands that a match criterion is arbitrary, that a perfect match is better than a bare match and that a near miss is not much different from a bare match. But a courtroom is not a scientific setting. Judges and juries balk when told that while the difference between the suspect and crime samples did not fall within the laboratory's definition of match, an expert believes it is a match. And it is reasonable for jurors to think that a match proportion calculated using a formal criterion is meaningless when an unspecified de facto match criterion, such as "visual match," is used to decide whether the suspect and crime samples match.

In contrast, the likelihood ratio approach I recommend is well adapted to making scientific inferences in courtroom settings. The likelihood ratio approach does not use an arbitrary match criterion. Rather, when assuming a normal error distribution, the likelihood ratio changes gradually, getting larger when the suspect's band weights are closer to the crime sample's. If suspect and crime sample band weights are moderately far apart on a particular probe, the likelihood ratio will be small, but it will not be zero. This small likelihood can be more than offset by likelihood ratios from other probes on which the bands are close together.

The likelihood ratio or Bayes factor is calculated as the ratio of the (integrated) likelihood of the evidence assuming guilt to the likelihood of the evidence assuming innocence; in particular, it does not depend on assumptions of prior probability of guilt. The posterior probability of guilt does depend on the prior probability of guilt. Only a juror (or the judge in nonjury cases) should assess the latter. But forensic laboratories might reasonably supply a table or graph demonstrating Bayes' theorem and showing how the genetic evidence serves to change prior probabilities into posterior probabilities.

The overwhelming proportion of criminal cases in the United States in which the prosecution uses DNA profiling have been successfully prosecuted. But in some cases laboratories have had a difficult time convincing the court of the appropriateness of their inferences. One such case is *New York v. Castro*, described in Sections 3 and 7. Another is *Maine v. McLeod.* The latter case centered on whether there was band shifting and how to adjust for it. The differences between the suspect's and crime sample's band weights were outside Lifecodes' matching criterion of ±3 s.d.'s for *all nine* bands of the four probes used. Lifecodes was still using the same s.d. as in the Castro case, and so their matching limits for the difference between two bands were ±1.8%. All nine differences were between 2.67% (or 4.45 s.d.'s) and 4.90% (or 8.17 s.d.'s), and in every case McLeod's band was *larger* than that of the crime sample. If the suspect and crime samples were from the same individual, then this is band shifting.

In *McLeod*, Lifecodes used a monomorphic probe called DXZ1 to correct for the alleged band shifting. (Recall from Section 3 that a monomorphic probe is the same in everyone, and so must be the same in the two samples.) McLeods' band on this monomorphic probe was 3.15% larger than the crime sample's, so Lifecodes subtracted 3.15 from the 9% differences of the bands on the other probes. The nine results were then between −0.48% and 1.75%—all in the ±1.8% range! The defense discovered that Lifecodes had used a second monomorphic probe, DYZ1, that they had not reported to the prosecution. McLeod's DYZ1 band was only 1.72% larger than the crime sample's. Subtracting this from the nine percentages left two of them outside the ±1.8% range! Hearing this ambiguity during a pretrial hearing in December 1989, the prosecution decided to drop the case.

A proper way to correct is to explicitly consider correlations among bands (including those on monomorphic probes when available) in formulating likelihoods. An analysis, such as that described by Berry, Evett and Pinchin (1991), gives a very large likelihood ratio in cases such as *McLeod*.

There are problems with using probabilistic arguments in court. Some courts allow them and some do not. (In Minnesota they are allowed in civil cases but not in criminal cases, although the Minnesota Supreme Court seems to be unique in this attitude.) Ellman and Kaye (1979) argue for their use. Fairley (1973) presents arguments on both sides. A major concern is the difficulty of communicating probabilities or probabilistic reasoning to jurors. For example, they may not interpret likelihood ratios of 100 billion and 100 any differently. (Actually, when the prior probability is greater than 50%, both give a posterior probability of greater than 99%, so maybe they shouldn't be interpreted very differently!)

In the disputed paternity setting, for each probe used there is only about half as much inferential ability as there is in cases of forensic identification: An individual has only one paternal gene. But using the combined evidence of several independent hypervariable loci can have enormous inferential power, even in a paternity case.

ACKNOWLEDGMENTS

Jon Seltzer did the computer programming that generated some of the tables and figures, and his comments and suggestions were invaluable in other aspects of this work. James Crow, Ian Evett, Peter Gill, David Gjertson, Eric Lander, Dennis Lindley and Ronald Pruitt made comments that were very helpful. Two anonymous referees made suggestions that greatly improved the paper. Michael Baird kindly provided raw data from some of Lifecodes' experiments. Research supported by in part by the National Science Foundation under Grants DMS-88-03087 and DMS-89-011548.

REFERENCES

Baird, M., Balazs, I., Guisti, A., Miyazaki, L., Nicholas, L., Wexler, K., Kanter, E., Glassberg, J., Allen, F., Rubenstein, P. and Sussman, L. (1986). Allele frequency distribution of two highly polymorphic DNA sequences in three ethnic groups and its application to the determination of paternity. *American Journal of Human Genetics* **39** 489-501.

Balazs, I., Baird, M., Clyne, M. and Meade, E. (1989). Human population genetic studies of five hypervariable DNA loci. *American Journal of Human Genetics* **44** 182-190.

Berger, J. O. (1986). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. Springer, New York.

Berry, D. A., Evett, I. W. and Pinchin, R. (1991). Statistical inference in crime investigations using DNA profiling. *J. Roy. Statist. Soc. Ser. C.* To appear.

Berry, D. A. and Geisser, S. (1986). Inference in cases of disputed paternity. In *Statistics and the Law* (M. H. DeGroot, S. E. Fienberg and J. B. Kadane, eds.) 353-382. Wiley, New York.

Cohen, J. E. (1990). DNA fingerprinting for forensic identification: Potential effects on data interpretation of subpopulation heterogeneity and band number variability. *American Journal of Human Genetics* **46** 358-368.

DeGroot, M. H. (1970). *Optimal Statistical Decisions.* McGraw-Hill, New York.

Ellman, I. M. and Kaye, D. H. (1979). Probabilities and proof: Can HLA and blood group testing prove paternity? *New York University Law Review* **54** 1131-1162.

Evett, I. W., Cage, P. E. and Aitken, C. G. G. (1987). Evaluation of the likelihood ratio for fibre transfer evidence in criminal cases. *J. Roy. Statist. Soc. Ser. C* **36** 174-180.

Evett, I. W., Werrett, D. J., Gill, P. and Buckleton, J. S. (1989). DNA fingerprinting on trial. *Nature* **340** 435.

Evett, I. W., Werrett, D. J., Pinchin, R. and Gill, P. (1990). Bayesian analysis of DNA single-locus profiles. In *Proceedings of the International Symposium on Human Identification* 77-101. Promega, Madison, Wis.

Fairley, W. B. (1973). Probabilistic analysis of identification evidence. *Journal of Legal Studies* **11** 493-513.

Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. In *Recent Advances in Statistics* (H. Rizvi, J. Rustagi and D. O. Siegmund, eds.) 287-302. Academic, New York.

Gjertson, D. W., Mickey, M. R., Hopfield, J., Takenouchi, T. and Terasaki, P. I. (1988). Calculation of probability of paternity using DNA sequences. *American Journal of Human Genetics* **43** 860-869.

Jeffreys, A. J., Wilson V. and Thein, S. L. (1985a). Hypervariable 'minisatellite' regions in human DNA. *Nature* **314** 67-73.

Jeffreys, A. J., Wilson, V. and Thein, S. L. (1985b). Individual-specific 'fingerprints' of human DNA. *Nature* **316** 76-79.

Kuo, L. (1986). Computations of mixtures of Dirichlet processes. *SIAM J. Sci. Statist. Comput.* **7** 60-71.

Lander, E. S. (1989a). DNA fingerprinting on trial. *Nature* **339** 501-505.

Lander, E. S. (1989b). Expert's report in People vs. Castro. Unpublished manuscript.

Lewin, R. (1989). DNA typing on the witness stand. *Science* **244** 1033-1035.

Lindley, D. V. (1977). A problem in forensic science. *Biometrika* **64** 207-213.

Lo, A. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *Ann. Statist.* **12** 351-357.

Morris, J. W., Sanda, A. I. and Glassberg, J. (1989). Biostatistical evaluation of evidence from continuous allele frequency distribution DNA probes in reference to disputed paternity and disputed identity. *Journal of Forensic Sciences* **34** 1311-1317.

Werrett, D. J., Gill, P. D., Evett, I. W., Lygo, J. E. and Sullivan, K. M. (1989). DNA analysis in Home Office Laboratories: Its introduction immediate future and statistical assessment. In *Proceedings of the International Symposium on the Forensic Aspects of DNA Analysis* 147-167. FBI Academy, Quantico, Va.