Comment

Donald A. Berry

Professor Roeder has presented the issues clearly, intelligently and in a balanced fashion. I agree with nearly everything she says. I will comment on some issues that she considers and address some she does not.

MATCH/BINNING VERSUS CONTINUOUS MEASUREMENT AND LIKELIHOOD RATIOS

I am happy to see Professor Roeder argue in favor of likelihood ratios tailored to the observed error distributions over a process initiated by declaring whether the suspect and crime samples match. I will say little about this issue here because, as she indicates, I have written about it elsewhere. The match/binning processes used by the FBI and U.S. forensic laboratories are consistent with the state of statistical knowledge before the time of Daniel Bernoulli. They are archaic and serve to handicap a fine technology. Some cases that might be won are lost, and otherwise good cases are not brought to court. The total number of such cases is small, but a few is a few too many. Fortunately, the technology is usually able to survive the primitive statistical methods.

This circumstance might have been avoided had statisticians been consulted during the development of DNA fingerprinting. There are now some excellent statisticians (such as Kathryn Roeder and Bruce Weir in the United States and Ian Evett in the United Kingdom) involved in the process, but momentum is difficult to interrupt (and statisticians are still not allowed in the inner sanctum—as Professor Roeder indicates, the 1992 NRC panel had no statisticians among its members). Instead of trying to patch up its inadequacies, the DNA fingerprinting community should start over from scratch, and this time with statisticians participating in formulating the statistics.

A further point about match/binning: I conclude from discussions with Bruce Budowle of the FBI that some match/binners still do not understand a very important tenet of inference. The matching proportion is a likelihood ratio. As such it can be used as a measure of evidence; but the definition of match that is applied to a suspect's profile must be applied as well to profiles in the database to determine a matching proportion. Otherwise the results are not interpretable. The FBI apparently uses different criteria to decide when a suspect matches than they use in calculating match proportions. I do not know whether they are more conservative in the latter than in the former: while they are quite open about the latter, I have never been able to pin them down concerning the former.

Donald A. Berry is Professor of Statistics at the Institute of Statistics and Decision Sciences, Duke University, Durham, North Carolina 27706.

INDEPENDENCE OF ALLELES

The controversy concerning independence of alleles has the wrong focus, as Professor Roeder suggests. The issue should not be whether alleles are independent, but what difference it makes to the conclusions. Usually, differences are minor. There may be correlations among the loci most commonly used, but none I have seen matter much. Some people would have us throw the baby out with the bath water.

In the previous paragraph I am not counting correlations among measurement errors (band shifting). These are very important. They are easy to handle using a continuous measurement, likelihood ratio approach, but not so easy using match/binning.

Assuming independence can make for extremely small match probabilities (or likelihood ratios). Tiny probabilities make me uneasy, as I will indicate below.

CONVEYING MATCH PROBABILITIES OR LIKELIHOOD RATIOS

According to an article in the *Minneapolis Star Tribune* (December 20, 1991): "If the two samples give the same results when tested, advocates of the [DNA fingerprinting] method say, it means the odds are overwhelming that one person was the source of both samples." This is reversing the conditional. It may be what "advocates of the method say," but it is most assuredly not what "it means." Statisticians have no trouble recognizing this, but almost every nonstatistician I have met has trouble understanding the difference between this and a correct state-

ment: "If the two samples give the same results when tested, this is overwhelmingly more likely when the samples are from one person than when there are two different sources."

This sentence is stilted, and I do not know how to make it more readable without making it wrong. Journalists hate stilted sentences. I know one—different from the one quoted above—who was willing to say that the likelihood ratio was the posterior odds even though I had convinced her this was incorrect. To my protest that she not say something she knew to be wrong, and after a long pause, she replied "Dr. Berry, my readers are scientists, not statisticians!"

Jurors are not statisticians either and so they must be confused by match probabilities. If jurors view them as posterior probabilities, it is difficult for them not to convict. If they are confused and appreciate that they are confused, they are most unlikely to be able to resolve their confusion without outside help, and they may react by discounting the evidence. Either result may lead to an injustice.

What is the resolution? I see no alternative to discussing Bayes' theorem. For example, one can present juries with curves or tables that relate prior and posterior probabilities in the case at hand. It is not important that jurors assess their prior probabilities, but it is important that they understand the need for such probabilities in properly interpreting evidence that is presented to them as match probabilities.

LABORATORY AND OTHER ERRORS

Some laboratories claim they do not make errors, and those that do not make such a claim act as though they do not. They do make errors. Only the frequencies by type of error are at issue. Is the rate one in a hundred? One in a thousand? Smaller yet? My credibility strains at rates smaller than one in a thousand, but even smaller error rates can greatly affect conclusions. The reason is that match proportions are usually tiny. A match probability of one in a billion is clearly an exaggeration in the presence of error rates on the order of one in a thousand.

Common practice among scientists is to ignore events of low probability; but when a conclusion is an extremely small probability, these events rise to the surface and become dominant. Human errors in mixing up or mislabeling samples must be a lot more common than one in a million. Perhaps it is my suspicious nature, but for any given crime I would put the probability that a laboratory technician is the true culprit at higher than one in a million. (I mean no malice toward laboratory workers, but a fiendishly cunning serial killer might seek out employment in a forensic laboratory where he could try to switch samples.)

To give another example, I am not a twin. I say this with confidence; but I would not bet my life on it. My probability that I am a twin is about one in a million. My probability that a suspect in a particular case is an identical twin is greater than that, even if he thinks he is not a twin; and the identical twin would match the crime sample whenever the suspect does. All these point toward adjusting very small match probabilities upward. (This of course can be done systematically using Bayes' theorem.)

The good news is that most people cannot fully appreciate tiny probabilities. For example, I doubt that a typical juror distinguishes between one in a million and one in a billion. Match probabilities as small as 1/738,000,000,000,000 have been reported. In view of my arguments above, this is at least nine orders of magnitude too small, but its impact on a jury may be the same as 1/1,000,000.

IDENTIFYING SUSPECTS BY SEARCHING THROUGH A DATABASE

I will address this issue in extended fashion, in part because it raises broader issues.

Suppose criminal investigators compare individuals in a database of DNA profiles with a profile obtained from a crime scene, and a case is brought to trial as a result. This does not fit the usual mode in which the DNA information is confirmatory, with the hypothesis of guilt having been generated in some other fashion. Clearly, a case is weaker if there is no other evidence. Professor Roeder indicates that the NRC panel recommends not using such information in court. The case is not that weak! Not using information is antithetical to scientific inference. (If nothing else, this should have been a clue to the NRC panel that they needed statistical help.) Why should DNA evidence be treated differently from eyewitness testimony, say? Eyewitnesses can testify that they picked the suspect out of a file of mug shots. Is eyewitness testimony more reliable than DNA fingerprinting? Just the opposite is true. Actually, eyewitness testimony is notoriously unreliable (Loftus, 1975).

Consider two circumstances. In the first, a particularly unreliable eyewitness picks Culbert Pritchard from a file of mug shots. Mr. Pritchard's DNA profile is found to match the evidentiary sample. In the second, Mr. Pritchard's DNA profile is contained in a database and is found as a result of a database search to match the evidentiary sample. Mr. Pritchard's photo turns out to be contained in a file of mug shots shown to the eyewitness, who picks out Mr. Pritchard. These two circumstances are identical except for order. The two pieces of evidence ought to be

treated symmetrically. Hypothesis generation and confirmation are treated asymmetrically in many applications of statistics—they should not be.

The inferential aspects of database searches are problematic for frequentists, but they are quite manageable from a Bayesian perspective. Suppose that (in the worse case) there is no evidence that points toward any particular individual over another in some population. So every individual has the same probability as every other. The critical issue is the denominator. Suppose the size of the population is N. Then the prior odds that any particular individual is the culprit equal $1/(N-1)$. The value of N will be subject to disagreement between the defense and the prosecution. It might be reasonable to take N to be the number of residents in a country—or males in the country if, say, the culprit is known to be male; but it is hard to imagine a crime for which it would be reasonable to take N to be the population of the Earth. In any case, the dependence on N of the posterior probability that an individual is the culprit can be examined.

A crime is committed in Denver. Take N to be the population of the United States, perhaps restricted to those capable of carrying out the crime (adult males perhaps). The likelihood ratio for the DNA evidence is calculated; call it LR(DNA). This is the probability of the DNA evidence given that Mr. Pritchard is the culprit divided by the probability of the DNA evidence given that he is not. The odds that Mr. Pritchard is the culprit then equal $\mathrm{LR(DNA)}/(N-1)$. Presenting LR(DNA) without multiplying it by $1/(N-1)$ would be misleading, and this is no doubt related to the concern of the NRC panel; but multiplication is a simple operation that puts LR(DNA) in perspective.

How to incorporate other evidence once Mr. Pritchard has been identified? This is tricky because of the uncertainties involved and because of the rules for conducting legal proceedings. Consider the answers to questions such as these: Who is this Mr. Pritchard? Is he a resident of Denver? Does he travel? Does he have an alibi? How reliable is it? Does he know the victim? Does he have a motive? Does he have a background consistent with the type of crime committed? Answering any of these questions gives a likelihood ratio: The probability of the answer given Mr. Pritchard is the culprit divided by the probability of the answer given that he is not. (The ratio is 1 if there is no answer.) Conditioning on the answer to question 1, the revised odds are $\mathrm{LR}(1) \times \mathrm{LR(DNA)}/(N-1)$—assuming the answer is independent of the DNA evidence. Then after question 2, $\mathrm{LR}(2) \times \mathrm{LR}(1) \times \mathrm{LR(DNA)}/(N-1)$ and so on. If there are dependencies then the appropriate probabilities are joint; suppose 1 and 2 are dependent but they are independent of the DNA evidence, the revised odds then equal $\mathrm{LR}(1,2) \times \mathrm{LR(DNA)}/(N-1)$.

For example, suppose Mr. Pritchard is found to have been in Denver at the time of the crime. The probability of this answer if he is the culprit is 1 (ignoring the possibility of error in the answer). The probability of the answer if he is not the culprit is the proportion of the population in Denver divided by the population of the United States. Alternatively and perhaps more appropriately for courtroom settings, an answer may serve to modify the value of N in the initial probability. For example, when it is found that Mr. Pritchard was indeed in Denver at the time of crime, N could have been taken to be the population of Denver at that time—giving the same answer as above, but skipping the multiplication.

Just because the answers to these types of questions are not generally quantified in court does not mean they cannot or should not be quantified in court. I do not know the admissibility of such quantitative answers, but it surely varies with the court. A prosecution might start with odds $1/(N-1)$ and then provide LR(DNA). The judge may or may not allow these two numbers to be multiplied. Then comes the answer to question 1 (Was Mr. Pritchard in Denver at the time of the crime?) with numerator and denominator of LR(1) calculated—most courts would allow these to be presented, or at least for question 1 to be addressed. Again, perhaps the division will be allowed and perhaps not, and perhaps providing $\mathrm{LR}(1) \times \mathrm{LR(DNA)}/(N-1)$ will be allowed and perhaps not. If calculations necessary for applying Bayes' theorem are not allowed, the prosecutors can rely on the jurors' internalized Bayes' theorems—in other words, the status quo! This resort is less than ideal because people's substitutes for Bayes' theorem are poor imitations, and errors can be in either direction.

I have glossed over the individuals in the database apart from Mr. Pritchard. I have assumed in effect that their average LR(DNA) is 1. This may not be so. The actual values of these LR's should be considered, and for two reasons. First, if their average is different from 1 this should be used to modify the above calculation. Each of the other $N-1$ individuals is assigned an LR, either calculated or not, with someone who is not in the database being assigned LR = 1. Then these $N-1$ LR's are added and the sum replaces $N-1$ in the above calculations. For example, should it turn out that another individual has the same LR as Mr. Pritchard, then the odds posterior to the DNA evidence that Mr. Pritchard is the culprit are no greater than 1 (posterior probability no greater than 1/2).

The second reason for considering the other LR's in the database is to check the assumptions of the DNA fingerprinting modeling process. For example, if the average LR is not close to 1, or if the distribution is different from the null distribution in an important

way, then this suggests a violation of the assumptions. It may be that the alleles were assumed to be independent and they are in fact highly correlated in the database being used. (There are many types of correlations. Many are irrelevant in a particular case. A correlation detected in looking at the distribution of LRs in the database is in a sense the only correlation that matters in the case at hand.) Or there may be an important degree of population substructuring. Indeed, a robust alternative to the usual process (robust at least for large databases) is to assume the average LR in the population is the average in the database and apply Bayes' theorem accordingly. For example, such an approach does not require assuming independence. The only concerns with this alternative are statistical: sampling variability, which is more of a problem for smaller databases, and whether the sample is random. I do not mean to minimize the latter concern: for example, population substructuring is still a problem.

EXPERTS AND ADVOCACY: BIASED TESTIMONY

Experts appear in court at the behest of the prosecution or defense. An attorney whose client has been charged in a violent crime in which DNA evidence is introduced wants to find scientists who will refute the evidence. An attorney for the prosecution wants to find scientists who will support the evidence. Sufficiently diligent searches will be successful. Opinions vary in every science. It is possible to find an expert with any given opinion who will testify [see Begley (1993) for examples]. There are various motivations for experts to testify. Some are laudable: science, common good, protecting individual rights. Some are not: money, notoriety, frequent flyer mileage. In any case, testimony of experts is biased. Judges and juries may know this and so discount expert testimony, but why should a legal system encourage testimony that ought to be discounted?

The searches mentioned above are eased by examining testimony in previous cases. As a result, the same experts tend to testify in case after case. In effect they become advocates, advocates for or against a technology. They—their persona, their thinking, their science—become objects of attack and to defend. Biases become more and more serious. Any semblance of objectivity disappears.

There must be a better way, one with less blatant bias. While this venue is hardly appropriate for recommending revisions in practices of jurisprudence, I will do so anyway: Expert testimony should be evaluated by expert witnesses supplied by the court. Court-appointed experts would be required to listen to the testimony of the prosecution and defense experts. Then they would express their reactions and opinions. They would not be paid but would be obliged to serve, with their service counting as jury duty in the municipality of their residence. If a municipality has no qualified experts, the court would import such from nearby municipalities and reimburse travel expenses. No system is bias-free, but this policy would rid the system of some extreme forms of bias.

Comment: Theory and Practice in DNA Fingerprinting

Richard Lempert

Throughout her useful paper on DNA identification, Professor Roeder properly attends to both theory and practice. Thus she acknowledges the theoretical soundness of certain criticisms that have been made of the standard paradigm used to evaluate DNA random match probabilities but argues that in practice these criticisms matter little. I am thinking here of the arguments that those cautioning against overweighing DNA evidence have made regarding the undeniable existence of population substructure and its potential implications for independence assumptions supporting the application of the product rule and for the use of convenience samples, such as data garnered from no more than a few local blood banks, to generate estimated allele frequencies for all Caucasians or African-Americans or Mexican-Americans living in the United States. Like Professor Roeder, I believe that these theoretically sound objections have, to date, been shown to be relatively unimportant in practice.

Richard Lempert is Francis A. Allen Collegiate Professor of Law and Professor of Sociology at the University of Michigan, Ann Arbor, Michigan 48109-1215. He is also acting chair of the Department of Sociology.