als to discussion of the frequency of the matching genotypes, the methods for computing that frequency and the controversy surrounding those methods may reinforce the powerfully prejudicial suggestion that false positives are a minor issue and that the frequency of the matching genotypes is the issue on which the value of DNA evidence will turn. In fact, where false positives are possible, the frequency of matching genotypes may have no relationship to the likelihood ratio that describes the value of the DNA evidence for proving two samples have a common source. Hence, it is at best an unhelpful statistic and at worse seriously misleading. Whether it should even be presented to juries is a question that I hope Kathryn Roeder, and her readers, will ponder.

Comment

B. S. Weir

Roeder has provided a useful review of the statistical issues involved in studies of human identification. She makes the distinction between objections to certain assumptions that might be raised in theory and the numerical consequences of those assumptions not being completely true in practice. A related issue is that of statisticians not taking into account all the relevant biological factors, and Roeder pointed to work of Geisser and Johnson (1992, 1993) in that context.

As Roeder explained, Geisser and Johnson explored the consequences of discretizing VNTR fragment lengths into a set of quantile bins, rather than the bins defined by viral fragment lengths as used by the FBI. Both binning strategies are adhoc, but the quantile bins lead to simpler analyses since each bin and each pair of bins is equally frequent. Roeder pointed out that the analyses of Geisser and Johnson have little relevance in the forensic debate since the problem of the unknown cause for single bands was ignored. The same point was made by Weir (1993), who also demonstrated that different numbers of bins, let alone different binning strategies, can lead to different conclusions regarding the independence of pairs of fragments in samples. The phenomenon has been well-documented in the population genetics literature.

Roeder herself might have referred to previous literature in her discussion of hierarchical Bayesian methods that invoke the Dirichlet distribution. Other authors have sought to use this distribution in the population genetics context (Rothman, Sing and Templeton, 1974; Spielman, Neel and Li, 1977), and there may be instances where it provides useful approximations. The current problem is to determine the conditional probability of a genotype, or VNTR profile, when that genotype has already been observed (for the perpetrator of a crime). Such conditional probabilities require the joint probabilities of *genotypes*, whereas Roeder in her equation (8) works with the joint probabilities of *alleles*. The joint genotypic frequencies require information about the relations between four alleles (two per genotype) rather than just two. Nichols and Balding (1991), in the paper that presented Roeder's equations (18), also ignored the relations between alleles considered three or four at a time. It is possible to approximate the necessary four-gene measures of identity with the two-gene measure called θ_s by Roeder, and θ or F_{ST} by others (Weir, 1994).

A deeper question concerns estimation procedures for θ. This quantity provides the correlation for alleles within the same subpopulation, and consequently it provides the component of variance between subpopulations in an analysis-of-variance setting. Evidently such a parameter cannot be estimated from data in one subpopulation (e.g., Weir and Cockerham, 1984), or even from data from the whole population without knowledge of subpopulation structure. Apparently, Roeder et al. (1993) overcome this logical barrier in arriving at estimates by assuming a distribution for allele frequencies, in contrast to the approach of Cockerham (1969) that regards the true allele frequencies as unknown.

The problem with taking genotypic frequencies to have a Dirichlet distribution is that results contrary to genetic expectations can result. Jiang and Cockerham (1987) simulated populations subject to genetic drift and compared a moment estimator of θ derived from an analysis-of-variance viewpoint with

B. S. Weir is in the Program in Statistical Genetics of the Department of Statistics at North Carolina State University, Raleigh, North Carolina 27695-8203, and has the title of William Neal Reynolds Professor of Statistics and Genetics. He directs an NIH-funded Program Project in Statistical and Quantitative Genetics that supports theoretical and experimental research in the Departments of Statistics and Genetics.

the "RST" estimator that Rothman, Sing and Templeton (1974) derived from a Dirichlet model. The moment estimator was essentially unbiased for their parameter values whereas the RST estimator had about a 50% bias. The RST estimator had a standard deviation about half that of the moment estimator. Jiang and Cockerham concluded that the Dirichlet model performed poorly for the genetic drift process, and were concerned that the model may not be broadly useful.

Notwithstanding these comments, the paper by Roeder is a welcome addition to the literature. It illustrates the role statisticians have to play in addressing societal issues.

Rejoinder

Kathryn Roeder

I would like to thank the discussants for their lively remarks, even those wide of the mark. Because of the subject of my review, I am not surprised by some of the emotional arguments put forth, although they seem out of place in *Statistical Science*. As Professor Lempert comments, "there is a kind of passion to each side, which sometimes seems, however politely, to amount to questioning the bona fides of the other." Before discussing the commentators' remarks in detail, I will outline their points.

The discussants broach several interesting issues that are far afield from the points covered in my review. The statistical issues in human population genetics, the core of my review, have been the focus of controversy in the courts and the scientific literature for the last few years. Professors Berry, Lempert and Weir agree with me that the criticisms leveled at the standard paradigm for estimating DNA profile probabilities, while sometimes sound in theory, have a negligible impact on the calculations in practice. Professors Balding, Donnelly and Nichols (BDN) and Professor Lewontin continue to question some population genetic assumptions upon which the probability calculations are based. Professor Sudbury stands alone in questioning the need for the paradigm. The adequacy of the genetic model and the importance of the choice of reference population are elaborated in Sections 1 and 4 of my rejoinder.

Several commentators raise concerns about laboratory error, something I did not discuss in depth in my original paper. They worry that samples will be mixed up in the laboratory, resulting in the suspect's sample being compared with itself, rather than with the crime sample. Another concern they raise is cross-contamination, which could also lead to an erroneous match. Professor Lewontin says that the danger of this is greater when a molecular technique known as PCR is used. I think that the danger of error depends more strongly on laboratory protocol than on the molecular technique.

From his comments, it seems that Professor Lewontin is unfamiliar with the protocol and methodology generally used by forensic scientists. He asserts that crime scene samples, being of limited quantity, are amplified using PCR. In fact, PCR is generally not used for the purpose he describes, and the genetic evidence presented at trial is usually not the product of PCR amplification. The major forensic testing laboratories (FBI, Lifecodes and Cellmark) do not regularly use PCR now, let alone in the past [Ivan Balazs, Director of Research at Lifecodes, and Bruce Budowle, Director of Research at the FBI Laboratory (Balazs and Budowle, 1993)]. Although PCR is sometimes used for an initial screening, for the bulk of cases forensic testing laboratories ultimately use a less sensitive method called RFLP typing via Southern blotting (NRC, 1992). Perhaps Lewontin's remarks are aimed at what he envisions for the future. Indeed RFLP analysis will eventually be replaced by some amplification process because results for the latter can be obtained almost immediately, whereas results for any RFLP analysis require four to six weeks or more.

Professors Thompson, Lempert and Berry believe that the average probability of a laboratory error should place a lower bound on the probability of a match. I disagree. A case-specific, posterior probability of a laboratory error is the appropriate calculation. Such a calculation, if admissible in court, should be presented separately from the probability of a match. Relying on the NRC report, Professor Thompson argues that the probability of a laboratory error should be estimated using proficiency testing. From the statistical perspective, it is clear that proficiency testing is not an efficient means of estimating a small probability. In Section 5, I discuss laboratory error in general, including methods of estimating the probability of error.

BDN voice concern about likelihood ratio statistics that calculate the probability of a match between un-