For $n = 16$, the maxima are $K = 4.90$ and $K_1 + K_2 = 4.51$. Thus, there is only a slight loss in information in using two independent loci.

To summarize verbally the above mathematics, a locus with n alleles is maximally informative when all alleles are equally frequent in the underlying population. The condition of equally frequent alleles is admittedly extreme, but one might approximate it by the appropriate choice of test loci. Under the assumption of maximum information, two loci with $\sqrt{n}$ alleles each are jointly about as informative as a single locus with n alleles. If the more informative single locus suffers from band overlap, its information content is diminished. One can take the approach of Berry and try to extract the most information by dealing with the quantitative measures directly. By comparison, the FBI's method of preset bins loses some information in discretizing the problem. According to the above mathematical argument, these bins should be equally probable rather than equally spaced.

In closing, let me stress that the rapid rate of innovation in molecular genetics is apt to overcome the technical problems such as band overlap associated with the current DNA fingerprint loci. These loci are typed by a procedure called Southern blotting. The alternative PCR techniques advocated by Weber and May (1989) and Budowle, Chakraborty, Giusti, Eisenberg and Allen (1991) are exquisitely sensitive to minute amounts of DNA and can avoid the problem of band overlap. However, PCR sensitivity can be so extreme that contamination by exogeneous DNA is troublesome. It is not now clear which technology will prevail. I prefer the PCR technology since it permits more loci to be typed from a small crime sample. As I have attempted to argue, most of the controversies over Hardy–Weinberg equilibrium and, particularly, linkage equilibrium will dissipate with better defined loci. In any event, we should welcome the inevitable improvements in technology even if the statistical issues become less interesting. Justice will be better served by greater genetic clarity.

ACKNOWLEDGMENTS

Research support in part by the University of California, Los Angeles; Harvard University; and USPHS Grant CA 16042.

Comment

Herman Chernoff

1. INTRODUCTION

Berry sets two objectives in his abstract. One is to introduce the Bayesian approach to the forensic use of DNA evidence, and the other is to compare that approach with that of "match/binning." The latter is criticized as giving results that are too extreme and for failing to distinguish, in principle, between results that barely fail to fall in the appropriate bin and those that are way out. As he points out, two potential observations that are very close to one another could lead to drastically different conclusions. As I understand it, he seems to suggest that the users of this approach may have recognized this problem in the Castro case and reacted by an ex post facto widening of the bin when the observation barely failed to fall in a bin suggesting guilt.

In my opinion the Bayesian approach is well suited for this subject and deserves to be developed as a useful tool. This approach has several difficulties, some of which are addressed by Berry. One of these is that of educating the members of the legal system and the potential jurors. Another is the use of density estimation to determine the frequency distribution of band weights.

Several issues will be discussed here. The match/binning approach, as described, doesn't make much inferential sense, and if the Bayesian approach should be compared with something, it should be with a more or less classical significance or frequentist Neyman–Pearson (NP) competitor. While binning does replace a continuous analysis by a discrete analysis that leads to aggravating discontinuities, one ought to evaluate the resulting cost in loss of efficiency before outlawing the practice of binning. To do so, we shall review briefly

Herman Chernoff is Professor of Statistics, Harvard University, Science Center, 1 Oxford Street, Cambridge, Massachusetts 02138.

some of the relevant, although asymptotic, properties of Kullback-Leibler (KL) information. These are relevant also to the argument, made in Lange's discussion, for the use of less polymorphic probes in forensic DNA analysis. Finally we shall comment on the handling of the problem of density estimation, the choice of the reference population, conditioning and the distinction between matching and guilt.

2. SIGNIFICANCE APPROACH

One basic flaw in the match/binning approach is that of multiplying probabilities. Aside from the possible lack of independence, there is a misleading implicit interpretation. If we have three independent pieces of evidence, with probabilities of 0.1, 0.2 and 0.3, under the hypothesis of innocence, the probability of the combination is 0.006. That probability could be reduced to an arbitrarily small number by adding additional pieces of "evidence" with no relevance at all to the issue of guilt or innocence. The resulting numbers have no inferential meaning.

The classical approach is to use significance tests where one calculates the P values or probability of achieving evidence, as measured by the value of a test statistic, as strong or stronger than that observed. It isn't clear from Berry's text whether the Lifecodes probabilities were P values. The numbers he used were apparently P values based on the test statistics corresponding to falling or not falling in an appropriate bin. However they were derived, the P values for the separate probes should not be multiplied. One standard approach to combining P values is to note that if the hypothesis being tested is true and the P value has a continuous distribution, then that distribution is uniform and its negative logarithm, which can serve as a test statistic, has the exponential or $\Gamma(1, 1)$ distribution. One natural approach to combining n P values is to use, as a test statistic, the sum of these negative logarithms which has a $\Gamma(1, n)$ distribution, i.e., the distribution of half of the chi-square random variable with $2n$ degrees of freedom, when the P values are continuously distributed. The resulting P value is based on the product of the individual P values but is larger, depending on n.

That approach is reasonable if the various P values being combined correspond to comparably powerful independent tests. As we shall point out later, if the P values derive from tests with wildly divergent powers, some adjustment would be in order. Note that the presence of some experiments that are not informative would tend to degrade the apparent significance if the sum of the negative logarithms is used and the hypothesis is false. An alternative is tentatively proposed in the next section.

3. KULLBACK - LEIBLER INFORMATION

The KL information for testing a simple null hypothesis $H_0 : \theta = \theta_0$ against a simple alternative $H_1 : \theta = \theta_1$ is defined by the equation

$$K = -E_{\theta_0}\left\{ \log \frac{f(X, \theta_1)}{f(X, \theta_0)} \right\}.$$

It has several relevant asymptotic properties for likelihood-ratio tests based on n independent observations (see Chernoff, 1979). If the type 1 error, i.e., the probability of rejecting the hypothesis H_0 when it is true, is kept fixed, the other error probability approaches 0 at the exponential rate in n given by K. Related to this is the fact that the P value for testing H_1 versus H_0 will tend to be roughly of the order of $\exp(-nK)$ when H_0 is true. This asymptotic result can be extended to independent observations with different distributions, corresponding to the use of different independent probes, by applying the additivity of the KL numbers for combining different experiments. Thus the use of KL numbers is of value in determining what are good probes to use, i.e., in deciding how to design experiments.

On the other hand, it is important to bear in mind that KL refers to an asymptotic result and should be applied with some reservations. For example, two experiments with $K = \infty$ are not necessarily equivalent, nor are they necessarily better than one with a finite value of K in the finite sample situation. If one tests the hypothesis that an unknown probability p is 0 against the alternative that it is some specified value p^*, the infinite value in this case reflects the fact that, if p is not zero, that will be determined with certainty in some finite number of trails when an event finally takes place. In this case, the power of the test will obviously depend on the value p^*. Furthermore, if one were confined to only one observation, I might prefer to test $p = 0.2$ versus $p = 0.8$ with finite K rather than test $p = 0$ versus $p = 0.01$ with infinite K. Nevertheless the value of K is a useful measure for discriminating between experimental setups, and Lange uses it to argue in favor of the use of less polymorphic probes. It was introduced here to provide a quantitative evaluation of the practice of binning, so that we need not depend on vague qualitative arguments in an area where technology is bound to change rapidly the parameters of the discussion. In extreme cases such as the

one illustrated, a better understanding of the value of the experiment comes from a study of the probability distributions of the likelihood-ratio under both of the relevant hypotheses.

Suppose that the hypothesis H_M states that there is matching on each of k separate items and the alternative is that, for each of these, the two variables being observed are independently selected from the same reference population. Then some theoretical considerations, on which I won't elaborate, suggest the use of the following test statistic based on the individual P values and the corresponding KL numbers

$$T = -\frac{\sum K_i^{1/2}\Phi^{-1}(P_i)}{(\sum K_i)^{1/2}},$$

where Φ is the standard normal c.d.f. and K_i is the KL number for the ith item.

Assuming continuity of the distributions of the individual P values, this test statistic is normally distributed with the mean 0 and variance one if the hypothesis is true. My inclination is to stick with the simpler test suggested previously unless there is a considerable mismatch in the powers of the individual tests.

4. THE EFFICIENCY OF BINNING

Two types of binning were described. The kind practiced by the FBI is essentially one similar to a crude rounding off of the data. There does not seem much point to it in our context, but it may make sense as a data compression device if it is part of a practice being used for other purposes or if there are some other practical considerations involved, which are not discussed by Berry. The method used by Lifecodes is substantially different.

To discuss the efficiency of the various forms of binning, let us consider two hypotheses. The matching hypothesis, H_M, typically associated with guilt, is that two variables X_1 and X_2 are noisy observations on a common random variable, randomly selected from some reference population. The nonmatching hypothesis H_N is that both X_1 and X_2 are independently distributed with the same marginal distribution as under H_M. Then the KL information is

$$K(H_M, H_N) = E_{H_M}\left\{\log\frac{f_2(X_1, X_2)}{f(X_1)f(X_2)}\right\},$$

where f_2 is the joint density under H_M and f is the corresponding marginal.

If X_1 and X_2 have a bivariate normal distribution with correlation coefficient ρ, then $K = -0.5\log(1-\rho^2)$. If we replace X by X^* representing one of the $2k$ bins with end points μ, $\mu \pm \delta\sigma$, $\mu \pm 2\delta\sigma, \ldots, \mu \pm (k-1)\delta\sigma, \pm\infty$, we achieve values of K depending on k, δ, and ρ. By selecting δ so that $(k-1)\delta = 3$, we can tabulate the corresponding values of K, in Table 1. The use of continuous data is referred to by $k = \infty$.

In the method used by Lifecodes, the bins are not applied to X_1 and X_2, but to $|X_1 - X_2|$, yielding a 0 or 1 value representing in or out of the bin. Presumably, Lifecodes planned to use these bins in an exclusionary way, so that any probe which leads to an "out" value would determine nonmatching (innocence) for the whole case. Let us be less absolute about it, and we must be so if we use several probes with a large relative error in measurement. Then the result of the bin is a one or zero with the probability of a zero being p_M or p_N, depending on which hypothesis is correct. Our test criterion may require more than one "out" value or 1 to be observed to declare nonmatching or innocence if many probes are used. Here, we have for a given probe,

$$K = p_M \log\frac{p_M}{p_N} + (1-p_M)\log\frac{1-p_M}{1-p_N}.$$

The Lifecodes method is basically a crude binning method and it is no surprise that the ensuing loss of efficiency can be substantial. Table 2 presents the optimal values of δ and the resulting KL numbers, for the various values of the correlation ρ, assuming the bivariate normal model, and using the threshold $\delta\sigma_{X_1-X_2}$. The loss of efficiency compared to the use of continuous data ($k = \infty$ in Table 1), is relatively small when the correlation is large. In contrast, the FBI type of bins degrade relative to $k = \infty$ when ρ is large.

It would be more appropriate to use a model with a normal noise imposed on the histogram for D17S79 exhibited in Berry. Note that for that case, with the model $X = Y + Z$ where Z is the error in measurement with standard deviation of about 0.0168 and the standard deviation of Y is about 0.34, the value of ρ is 0.9976. Here the unbinned

TABLE 1
Kullback-Liebler information for bivariate normal with correlation ρ using $2k$ bins and continuous data ($k = \infty$)

$k \setminus \rho$	0.00	0.40	0.80	0.90	0.95	0.99	0.999
2	0.00	0.04	0.19	0.29	0.37	0.52	0.64
3	0.00	0.06	0.30	0.45	0.57	0.80	0.98
4	0.00	0.07	0.39	0.59	0.75	1.04	1.28
5	0.00	0.08	0.44	0.67	0.87	1.21	1.51
10	0.00	0.09	0.50	0.79	1.08	1.64	2.10
∞	0.00	0.09	0.51	0.83	1.16	1.96	3.11

TABLE 2
Optimal values of δ and KL using bins $|X_1 - X_2| \le \delta\sigma_{X_1-X_2}$ for the bivariate normal with correlation ρ

ρ	0.40	0.80	0.90	0.95	0.99	0.999
δ	1.51	1.65	1.75	1.84	2.02	2.18
KL	0.005	0.14	0.33	0.57	1.26	2.35

KL number is 2.52, which differs little from the normal model value of 2.67. I was surprised, for I would have expected a somewhat larger value for the bumpy distribution.

5. DENSITY ESTIMATION

Because the Bayesian approach does not provide a simple estimate of the distribution of the data for the reference population corresponding to one of the probes, Berry uses a simple kernel estimator, which tends to smooth the observed distribution. One excuse for using the smoothing estimator is the danger of overreacting to an observation that occurs in a region that is relatively empty because the density is low there and our sample size is not infinite. But that treatment does not address the problem directly. The fact is that the observed density is the convolution of the "true" density, which is probably quite discrete and bumpy, with the distribution of the noise due to the measurement error, and is already smoothed. What is called for in the density estimation is not a smoother, but a deconvolution operation that would operate in the opposite fashion. In fact, the use of a smoother tends to put the accused at greater risk, because it is more likely that the specimens will be at locations of high density, which will be reduced in the estimation by a smoother. In that case, the evidence will seem stronger for matching than it should.

The possibility that we are dealing with a location that has positive probability but is not represented in the sample is a real problem. But that could be faced more directly by the technology of coverage probability (see Robbins, 1968). According to that theory, the total population probability allocated to intervals that have no observations can be estimated by the proportion of observations which fall into singleton cells, i.e., cells which have only one observation, because both of these variables have the same expectation. Thus, in the case of D2S44, there are 13 singletons out of 295 observations and hence about 4.5% of the total population for the (convolved) distribution has yet to be represented. This gives an estimated upper bound to how much should be allocated in the equation for H_2 in Berry.

The 4.5% seems rather high and conservative. One can reduce that estimate by using the number of doubletons, because the sum of the squares of the probabilities for the underrepresented cells is estimated by twice the number of doubletons divided by the square of the number of observations. In this case, there are seven doubletons and the estimate is 0.00016, which suggests that the largest probability for an unrepresented cell would not be any greater than 0.013. Without pursuing the issues of accuracy and methods of deconvolution further, it is reasonable to expect that these represent a more direct attack on the issues than does the ad hoc smoothing techniques suggested by Berry.

6. MISCELLANEOUS

We conclude with brief mention of three miscellaneous issues.

Reference Population. The choice of the reference population raises some puzzling issues. Berry states that one should use that of the criminal, but generally the criminal is not known. Usually it would make sense to use the population to which the suspect belongs, if the issue is whether the DNA belongs to the suspect. Using the suspect's reference population would tend on the average to make the evidence seem weaker than if some other population were used and would lead to conservative results favoring the suspect. That is only fair for if the suspect were guilty this is the correct reference population, but if he were innocent, one would wish to be as conservative as possible.

In the Ponce case the issue was whether the blood belonged to the victim, and there the natural reference population is that of the victim.

The above discussion has assumed implicitly that there is a natural reference population for each individual. That is questionable. Should one lump all Hispanics together, or separate Cubans from Puerto Ricans and from Argentineans, etc.? If we do this, we start to lose the sample size necessary to estimate the distribution of band weights. I would conjecture that pragmatic answers to this question might be easier to obtain if probes with few alleles were used. In any case, this represents a theoretical issue that is worth exploring.

Guilt versus Matching. Several hypotheses have to be distinguished in a careful discussion. These are Guilt, Matching of the DNA, and Matching of Band Weights for several probes. (For convenience in discussion, Berry has identified these hypotheses.) Matching of the DNA strands may not imply guilt. Even if Ponce's blood were on Castro's watch, it would not, by itself, constitute proof positive of

guilt, although it would be very strong evidence. If there were identical twins involved, or even siblings, the force of evidence might be reduced. Matching of band weights for several probes does not imply that the entire DNA system is the same. In fact, if only a few probes with few alleles matched, the evidence could be far from overwhelming, even in those cases where there is no error in measurement.

It should be remarked that if Lifecode type bins are used in a nonexclusionary fashion, apparent matching on almost all of the highly polymorphic probes could be strong evidence, even if there were apparent failure to match on one or two probes. The strength of the evidence would depend on the bin sizes and the relative magnitude of the errors of measurement.

Conditioning. One advantage of Bayesian analysis over NP analysis is that, in the former, we can examine the evidence as it arrives. A partial reply involves the use of conditioning. The force of classical inference is sometimes strengthened by using conditioning appropriately. That could be done here, too. For example, once we had measurements on Ponce's blood, we could use those data to help select what constitute effective probes for the comparison.

ACKNOWLEDGMENT

This research was supported in part by NSF Grant DMS-88-17204.

Comment: Uncertainty in DNA Profile Evidence

D. H. Kaye

Donald Berry's article on inferring identity from DNA profiles presents a method for "direct calculation of the probability that the suspect is guilty" and "the probability that an alleged father of a child is the true father." The method is Bayesian. Berry computes the posterior odds of guilt as the product of the prior odds (assessed on the basis of all the evidence apart from the electrophoretic measurements) and a likelihood ratio for the DNA results. The specific likelihood ratio (R) that he skillfully derives for single-locus restriction fragment length polymorphisms accounts for random measurement error in electrophoresis and for sampling error in a laboratory's data base on the distribution of fragment lengths in the population.

This comment examines, from the perspective of a lawyer, two connected issues: the forensic importance of quantifying measurement and sampling error and the desirability of combining likelihoods and priors for jurors or judges. I try to place Berry's treatment of these matters in the context of the emerging case law on DNA profiling, and I speculate about the advisability of bringing Bayes to the bar. Proceeding on the premise that Berry's mathematics is impeccable, I conclude that if "To bin or not to bin?" is the question, then Berry has the answer.

D. H. Kaye is Regents Professor, Arizona State University, College of Law, Tempe, Arizona 85287-7906.

1. LABORATORY MEASUREMENT ERROR

Berry's analysis handles normally distributed (and log normal) laboratory errors in measuring the position of a perceived band, and he notes that a more complex analysis could handle other continuous error distributions. Yet, much of the criticism of forensic DNA work emphasizes other threats, such as contamination and degradation of samples. Thompson and Ford (1991, page 138), for example, report that missing bands, extra bands and systematically shifted bands are "quite common in the forensic casework."

These types of experimental error have received considerable judicial attention. Virtually all courts in the United States to face the issue have held that DNA findings of identity are potentially admissible, but the degree of experimental rigor actually required for admission varies with the understanding of the court and the persuasiveness of the experts. *People v. Castro* (1989) is remarkable for the extent of the judicial inquiry into