VOLUME 83 JANUARY 1970 NUMBER 3

HARVARD LAW REVIEW

A BAYESIAN APPROACH TO IDENTIFICATION EVIDENCE

Michael O. Finkelstein * *and William B. Fairley* **

State courts have met little success in analyzing whether prosecutors should be permitted to introduce mathematical statistics to show the weight that should be given to identification evidence. Taking one case as a convenient paradigm, the authors demonstrate that statistics will rarely conclusively identify a defendant. They suggest a mathematical method appropriate where statistical evidence alone is inconclusive. When other incriminating evidence raises a suspicion apart from the statistical evidence, Bayes' theorem can be applied to indicate the degree that the inconclusive statistical evidence heightens the suspicion.

IN *People v. Collins*,[1] the Supreme Court of California rejected a prosecutor's effort to link the defendants to a crime by using mathematical statistics. The decision is significant because the judges took the prosecutor's statistical sortie seriously enough to comment at length on the problem of statistical proof and to attempt a mathematical demonstration of the correct form for such analysis. In *Collins*, both the accused and the apparently guilty pair were interracial couples. The same mathematical approach was used in earlier cases to help make identifications (more plausibly) on the basis of similarities in typewriting, handwriting, fibers, or hairs.[2] Because of the development of new

* Member of the New York Bar. Lecturer in Law, Columbia University. A.B., Harvard, 1955; LL.B., 1958.

** Assistant Professor of Statistics, New York University Graduate School of Business Administration. A.B., Swarthmore, 1960; Ph.D., Statistics, Harvard, 1968.

We wish to thank Kenneth Jones of Columbia Law School and Stephan Fienberg of the Department of Statistics, University of Chicago, for their readings of the manuscript.

[1] 68 Cal. 2d 319, 438 P.2d 33, 66 Cal. Rptr. 497 (1968) (en banc).

[2] The principal reported cases in which such statistical evidence has been presented are Miller v. State, 240 Ark. 340, 399 S.W.2d 268 (1966); People v. Jordan, 45 Cal. 2d 697, 290 P.2d 484 (1955); People v. Trujillo, 32 Cal. 2d 105, 194 P.2d 681, *cert. denied*, 335 U.S. 887 (1948); State v. Sneed, 76 N.M. 349, 414 P.2d 858 (1966); People v. Risley, 214 N.Y. 75, 108 N.E. 200 (1915). *See The Howland Will Case*, 4 AM. L. REV. 625 (1870), *discussing* Robinson v. Mandell, 20 Fed. Cas. 1027 (No. 11959) (C.C.D. Mass. 1868).

techniques for analyzing the composition of small fragments,[3] evidence of this latter sort, backed with statistics, is likely to appear more frequently in future court proceedings.

This article discusses the use of mathematics to appraise the significance of such evidence. It is our general conclusion that the approach taken in cases such as *Collins* ought to be abandoned because it is appropriate to extremely few situations, and those can be handled without statistical analysis. Moreover, the alternative proposed in the *Collins* appendix was not correct in that case and will not generally be useful. We propose a new approach, based on Bayesian probability analysis.[4]

I.

In *Collins*, an elderly woman walking home in an alley in the San Pedro area of Los Angeles was assaulted from behind and robbed. The victim said that she managed to see a young woman with blond hair run from the scene. Another witness said that a Caucasian woman with dark blond hair and a ponytail ran out of the alley and entered a yellow automobile driven by a male Negro with a mustache and beard. A few days later officers investigating the robbery arrested a couple on the strength of these descriptions,[5] and charged them with the crime. At their trial, the prose-

---

[3] *See, e.g.*, R. COLEMAN, F. CRIPPS, A. STIMSON, & H. SCOTT, THE DETERMINATION OF TRACE ELEMENTS IN HUMAN HAIR BY NEUTRON ACTIVATION AND THE APPLICATION TO FORENSIC SCIENCE (U.K. Atomic Energy Auth., Atomic Weapons Research Establishment Report No. o-86/66, 1967).

[4] Thomas Bayes' AN ESSAY TOWARDS SOLVING A PROBLEM IN THE DOCTRINE OF CHANCES (1763) is generally regarded as the first work to develop this method. A facsimile of Bayes' paper has been published under the direction of W. Edwards Deming (1963).

For an abstract discussion of the possibility of using Bayesian theory in trials, see Kaplan, *Decision Theory and the Factfinding Process*, 20 STAN. L. REV. 1065, 1084 (1968).

For a non-mathematical discussion of the traditional, non-Bayesian approach, see Note, 1967 DUKE L.J. 665. Mathematical treatments may be found in Kingston, *Probability and Legal Proceedings*, 57 J. CRIM. L.C. & P.S. 93 (1966); Kingston, *Application of Probability Theory in Criminalistics*, 60 J. AM. STAT. ASS'N 70, 1028 (1965); Kingston & Kirk, *The Use of Statistics in Criminalistics*, 55 J. CRIM. L.C. & P.S. 514 (1964); Mode, *Probability and Criminalistics*, 58 J. AM. STAT. ASS'N 628 (1963).

A Bayesian approach to the judicial determination of paternity has been proposed. H. STEINHAUS, THE ESTABLISHMENT OF PATERNITY, PRACE WROCLAWSKIEGO TOWARZYSTWA NAUKOWEGO ser. A., No. 32, at 5 (1954); Lukaszewicz, *O Dochodzeniu Ojcostwa* (*On Proving Paternity*), 2 ZASTOSOWANIA MATEMATYKI (APPLICATIONS OF MATHEMATICS) 349 (1955) (discussed at pp. 505–09 *infra*).

[5] When defendants were arrested the woman's hair was light, not dark blond, and the man did not have a beard. There was some evidence that the man had altered his appearance after the date on which the offense had been committed.

cution called an instructor of mathematics at a state college in an attempt to establish that, assuming the robbery was committed by a Caucasian blond with a ponytail who left the scene in a yellow car accompanied by a Negro with a beard and mustache, the probability was overwhelming that the accused were guilty because they answered to this unusual description. The witness testified to the "product rule" of elementary probability theory. This rule states that the probability of the joint occurrence of a number of mutually independent events equals the product of the individual probabilities of each of the events. The prosecutor then had the witness assume the following individual probabilities of the relevant characteristics:

Yellow automobile	1/10
Man with mustache	1/4
Girl with ponytail	1/10
Girl with blond hair	1/3
Negro man with beard	1/10
Interracial couple in car	1/1000

Applying the product rule to the assumed values, the prosecutor concluded that there would be but one chance in twelve million that a couple selected at random would possess the incriminating characteristics.[6] The jury convicted. On appeal, the Supreme Court of California reversed, holding that the trial court should not have admitted the evidence pertaining to the mathematical theory of probability.

The Supreme Court objected to the expert's testimony on several grounds. First, the record was devoid of evidence to support any of the six assumed individual probabilities. This objection is clearly justified. Some evidence of those probabilities is surely required as a foundation for such testimony. However, evidence sufficient to support a finding that the probability estimates are likely to be greater than the true values should suffice. This is significant because it may often be possible to justify generous estimates of probabilities which cannot be determined exactly.[7]

Second, the court found no proof that the six factors were statistically independent. Again the court was correct. If traits are not independent, but rather tend to occur together, then the multiplication of the individual probabilities of each factor usu-

---

68 Cal. 2d at 323 n.5, 438 P.2d at 35 n.5, 66 Cal. Rptr. at 499 n.5. The car was only part yellow. *Id.* at 322 n.2, 438 P.2d at 34 n.2, 66 Cal. Rptr. at 498 n.2.

[6] The prosecutor gratuitously added his estimation that the "chances of anyone else besides these defendants being there . . . having every similarity . . . is somewhat like one in a billion." 68 Cal. 2d at 326, 438 P.2d at 37, 66 Cal. Rptr. at 501.

[7] See pp. 511–14 *infra*.

ally yields a composite probability that is far too small, even if the individual probabilities are accurate. For example, given the hypothetical probabilities in *Collins*, if every Negro man with a beard also had a mustache then the chance of a Negro man with a beard and mustache is one-tenth, not one-fortieth as indicated by the product rule.[8] Either the mathematical method must take correlations into account, or there must be sufficient evidence of independence of the factors.[9]

A first look at *Collins* thus reveals two requirements for the introduction of statistical analysis in evidence: the prosecutor must introduce evidence as to the probabilities of the individual factors and of the relations among them. The court also explored two obstacles to such proof. The first relates to the capacity of a jury to deal with statistical evidence, and will be discussed presently.[10] The second, as to which the court's analysis was wrong, cuts much deeper.

Writing for the court, Justice Sullivan asserted that "no mathematical equation can prove beyond a reasonable doubt . . . that only *one* couple possessing those distinctive characteristics could be found in the entire Los Angeles area."[11] He supported his conclusion with a mathematical demonstration purporting to show that even if a couple selected at random had only one chance in twelve million of bearing the incriminating characteristics, the expert witness could not conclude that the accused were probably guilty because it was quite possible (about a forty percent chance) that at least one other couple in the Los Angeles area had those same traits.

The court's argument is incorrect because the supporting mathematical demonstration was wrongly conceived. The court's proof begins with the probability of selecting a couple with the specified characteristics at random from the population. This is

---

[8] If Negro men with beards seldom have mustaches, the chance of a Negro man with both is smaller than one-fortieth.

[9] Other courts' assumptions of independence in cases like *Collins* have been deservedly criticized. *See, e.g.*, Kingston, *Probability and Legal Proceedings*, 57 J. CRIM. L.C. & P.S. 93, 94–95 (1966).

Whether the factors in *Collins* could, even theoretically, be independent depends on their interpretation. If the factor of "one-tenth Negro males with beards" means that one in ten Negro men has a beard, and the beard rate is the same for non-Negroes, the joint occurrence of this factor and the factors relating to the girl could possibly be independent of the factor "interracial couple in car." Conversely, if the beard factor is interpreted as a generous estimate that one man in ten is a Negro with a beard, and similarly for the factors relating to the girl, the joint occurrence of the man and girl factors would of necessity be highly correlated with "interracial couple in car."

[10] Pp. 495–96 *infra*.

[11] 68 Cal. 2d at 331, 438 P.2d at 40, 66 Cal. Rptr. at 504.

assumed, following the prosecution, to be one in twelve million. The court then proceeds to derive the probability that there are two or more such couples in the population. Because the court was dealing with an existing, finite population, the frequency with which couples with the identifying characteristics may be found in that population is identical to the probability of selecting one at random. Thus, the court's assumption that one in twelve million is a fair estimate of the probability of selecting such a couple at random necessarily implies that it is a fair estimate of the number of such couples in the population. The probability that couples with the fatal characteristics would appear more frequently could only have been determined by examining the precision of the estimate — an examination which neither the court nor the expert was able to make because the estimate was not the result of any statistically valid sampling procedure.[12]

---

[12] The formula derived in the court's appendix is

$$\frac{1 - (1 - Pr)^N - NPr(1 - Pr)^{N-1}}{1 - (1 - Pr)^N}$$

where Pr is the probability of selecting at random a couple with characteristics of the accused and N is the total number in the population. The court first assumed the total population of suspects to be twelve million and then showed (correctly) that its conclusion would not be affected if the population were assumed to be infinite. 68 Cal. 2d at 335, 438 P.2d at 42–43, 66 Cal. Rptr. at 506–07.

The court's formula generates the forty percent probability referred to in the text because it assumes a sampling of the population with replacement of the sampled couples, instead of sampling without replacement. The difference in result between these two methods frequently is not very great because the number sampled is small relative to the whole population. But in the experiment posited by the court the number of drawings for the sample is equal to the suspect population. In this circumstance the difference between replacement and nonreplacement is critical.

To see what the court's formula leads to, assume there are twelve million balls in an urn, each ball standing for a couple but only one (yellow) having the characteristics of the accused. The probability of selecting a yellow ball in a single draw from the urn is one in twelve million. A series of twelve million selections is now made; after each selection the ball is examined and thrown back into the urn from which it may be reselected. This series of twelve million selections is made repeatedly. The probability computed by the court is a fraction the numerator of which is the number of series in which two or more yellow balls were selected and the denominator of which is the number of series in which one or more yellow balls were selected.

This statistic obviously has nothing to do with the likelihood that a couple answering the description of the accused was correctly charged. For if there was only a single ball in the urn representing a couple with the characteristics of the accused, the court's formula would still yield a substantial probability of duplication (the same ball being picked twice) although by hypothesis the accusation was correctly made.

The method developed in the appendix is similar to the analysis in 50 MINN. L. REV. 745 (1966), a discussion of *Collins* published prior to the appellate decision.

The court's formula would have been relevant if it were assumed that nothing were known about the actual population of Los Angeles and the only available information concerned some unknown process by which it had been created. If the one-in-twelve-million figure represented the probability that a couple when created would have the fatal characteristics, then out of all possible populations of Los Angeles that could be produced by this unknown process, forty percent of those with at least one such couple would have at least two such couples.

The objection to this approach in *Collins* is that the one-in-twelve-million figure was intended by the prosecution and by the court to describe the actual population of Los Angeles and not as a parameter for a "generational" probability model. It is not valid to use as a generational probability an estimate intended to reflect the actual population, and then assume that since nothing was known about the actual population, the probabilities of various populations could be computed by calculating the hypothetical outcomes of the creation process. Moreover, a generational model will not usually be useful in the problems discussed in this article because in most cases it will be far easier to gain knowledge of the actual population by sampling than to define in probabilistic terms the forces producing it.

The statistical problem of the *Collins* case is that of estimating the very figure which the court took as its assumption, namely the probability that a couple selected at random would have the characteristics of the accused. That probability represents the frequency of couples meeting the description of the one placed at the crime. If a sufficiently precise estimate could be made that the frequency of such couples in the Los Angeles area was one in twelve million, it would be possible to state within reasonable margins for error that there was only one such couple in the Los Angeles area.

But as a practical matter the court was right to doubt that the prosecutor could show uniqueness. A derivation of such extraordinarily small probabilities with any useful degree of precision would be extremely difficult. In most cases, the estimate of the population frequency of evidentiary traces (of hair or incomplete fingerprints, for example) will have to be made on the basis of samples numbering at most a few thousand. As a result, probabilities of the magnitude involved in *Collins* would require an inference, based on a few thousand trials, that an event would occur once rather than more than once in millions of trials. Such an inference inevitably involves powerful assumptions which cannot be adequately supported without extensive data. Except in cases where the number of suspects is sharply

limited, it will almost never be practically possible to gather enough data to sustain a conclusion of uniqueness with any confidence.[13]

The approach in *Collins* thus makes the number of suspects critical. Determining this number, however, will usually involve wholly arbitrary decisions. Shall it include only those in the same neighborhood, the same county, the same city, state, or the entire country? The jury might be given a range of choices and the probability associated with each choice, but jurors cannot rationally choose when, as is usual, there is no evidence bearing on this issue. Setting a generous upper bound will usually defeat the proof: the incriminating characteristics will occur more than once in a sufficiently large population. Moreover, it is probably as difficult to decide intuitively how many "suspects" there are as to decide how many of the suspects have the incriminating characteristics.[14]

We now turn to the court's second objection to the use of statistics. The court reversed the Collins' conviction because it felt that the powerful statistics would cow a jury into overlooking the possibility that the basis for the calculations could be in error. The court was obviously right. However, correct statistical methods will usually have an effect opposite to that feared by the *Collins* court. Findings based on such statistics should generally weaken nonquantitative testimony based on the same evidence.[15] An expert's opinion that similarities between fragments (*e.g.*, of fingernails or hair) identify a defendant must rest on his limited experience with similar fragments. If to his knowledge no such similarities have been observed in fragments from different

[13] As we have calculated it, assuming independence, the probability estimates for the separate characteristics in *Collins* would have had to have been supported by a sample in the neighborhood of 400,000 in order to sustain the conclusion that there was only a small probability that the frequency of couples with the fatal characteristics in the population was two or more in twelve million.

[14] Another factor of considerable potential significance in this type of case (which the court did not discuss) is what can be called "selection effect." If there are, say, twenty characteristics or features which could be used for identification purposes, and the chance is one in a thousand that any given feature would match, the probability of one or more matches assuming innocence is approximately two of one hundred. A procedure by which the identifying feature is selected from a large group may thus critically affect the probabilities in these cases. *Cf.* People v. Trujillo, 32 Cal. 2d 105, 194 P.2d 681, *cert. denied*, 335 U.S. 887 (1948), where the expert examined a large number of fibers taken from the accused's clothing and from the scene of the crime and was able to make eleven matches. Applying the product rule he concluded that the probability was one in a billion that this many matches would have occurred by chance. A portion of the expert's testimony is reprinted in M. HOUTS, FROM EVIDENCE TO PROOF 325–29 (1956).

[15] The *Collins* court in fact reached such a conclusion, but, as we have seen, the method employed was erroneous.

sources, he may testify flatly that the two fragments have a common origin. But proper statistical methods, by invoking an experience larger than any expert's, may well yield an estimate that a fragment occurs several times in a large population, even though the expert would conclude there were no duplicates.[16] In addition, an expert witness may base his appraisal on a multitude of details imperfectly recognized and difficult to define or catalog — just as we know a face from a multitude of features. It is impossible statistically to take all such details into account. Statistical observation is of attributes that can be objectively measured; it cannot hope to have the richness of information involved in ordinary or educated recognition. For these reasons, the inference of identity from statistics will generally be weaker than expert judgment expressed in the usual way.

On its facts *Collins* was bizarre, and its pseudo-statistics scarcely can be taken seriously. But the method used in the case was entirely representative of more sophisticated efforts made in earlier cases in which the experts also applied the product rule to generate vanishingly small probabilities. The *Collins* court was right when it concluded that efforts to prove uniqueness usually will be futile. Few, if any, evidentiary traces can be demonstrated by statistical analysis to be unique to a defendant. There is, however, a class of traces, potentially useful as evidence, which could be shown to appear only infrequently, though not uniquely. What is the probative significance of such non-unique traces? We propose to show that non-unique traces generally deserve substantial evidentiary weight, and that by the explicit use of mathematical theory the data can be cast in a form permitting more effective use of this evidence by the jury.

II.

Let us suppose a woman's body is found in a ditch in an urban area. There is evidence that the deceased had a violent quarrel with her boyfriend the night before. He is known to have struck her on other occasions. Investigators find the murder weapon, a knife which has on the handle a latent palm print similar to defendant's print. The information in the print is limited so that an expert can say only that such prints appear in no more than one case in a thousand. We now ask the significance of this finding.

Under the approach taken in *Collins* there would be little probative value to the palm print evidence. If the number of

[16] These methods are described in part VI, pp. 511–14 *infra*.

potential suspects were as few as one hundred thousand, about one hundred persons would have such prints. This is hardly a unique event. And yet, intuitively, the finding of such a relatively rare print which matches defendant's is telling. After all, the prosecutor may correctly argue that defendant is a thousand times more likely to have committed the crime than someone selected at random from the population. Without the print evidence the case probably does not go to the jury. With it the jury probably convicts. The mathematical formulation in *Collins* thus seems grossly to understate the intuitive impact of this evidence.

The difference between the two formulations lies in the unexpressed premises behind them. Proof of uniqueness was demanded in *Collins* because it was assumed as a starting point for the mathematical analysis that defendants were no more likely to have committed the offense than anyone else in the "suspect" population. The same assumption in our hypothetical case implies that the print evidence merely places defendant among a group of one hundred persons any one of whom is equally likely to be guilty. The probability of defendant's guilt remains small, though increased a thousand-fold (from one in a hundred thousand to one in a hundred) by the print evidence.

The tacit assumption in *Collins* of no advance knowledge is inconsistent with the way we ordinarily view evidence. We tend to see a case as a whole; our appraisal of any bit of information depends on the rest of the testimony and our life experience.[17] Guilt is determined by a "cumulation of probabilities." [18] Slight additional evidence in support of an event about which we already have persuasive evidence is given considerable weight, while evidence which would otherwise be highly compelling is discounted if it does violence to our prior beliefs.[19]

When statistics are not involved, this cumulative perspective controls the probative significance of evidence.[20] The same per-

---

[17] As one court put it, "Every man's experience demonstrates that his beliefs are based upon a great number of circumstances . . . which, when combined together, give strength to each other . . . ." *Ex Parte* Jefferies, 7 Okla. Crim. 544, 551, 124 P. 924, 927 (1912).

[18] 1 F. WHARTON, EVIDENCE IN CRIMINAL CASES 8 (11th ed. 1935).

[19]

> Suppose a number of witnesses testify that they saw a man thrust his hand into a bucket of water, and on taking it out a hole remained in the water where the man's hand had been. It matters not how positive and direct such testimony was, no sane jury would accept it. Why? Because their past experience, based on circumstances, teaches them that it is contrary to the laws of nature . . . .

Ex Parte Jefferies, 7 Okla. Crim. 544, 546, 124 P. 924, 925 (1912).

[20] *See* People v. Trujillo, 32 Cal. 2d 105, 194 P.2d 681, *cert. denied*, 335 U.S. 887 (1948). For a discussion, see REPORT OF THE PRESIDENT'S COMMISSION ON THE ASSASSINATION OF PRESIDENT KENNEDY 124 (1964).

spective should be used when statistics are involved. In our hypothetical case, the analysis of the palm print evidence should begin with the fact that defendant was far more likely to be guilty than someone selected at random. Consistent with this approach, it has been said that statistical evidence of the kind we have been considering should normally not be sufficient to support an identification unless accompanied by other evidence that would form the basis for a "prior" estimate of identity.[21] This is an intuitive idea, but one that can be justified. We use Bayes' theorem for this purpose.

III.

We begin our discussion of Bayes' theorem by deriving an expression for the probability that defendant used the knife, assuming that an incriminating print from a right hand palm is found on it. In accordance with general practice, we denote this probability $P(G|H)$, where G is the event that defendant used the knife (or, as we shall say, that "there is identity" between defendant and the knife user) and H is the event that a palm print similar to defendant's is found. $P(H|G)$ is the probability of finding the print assuming there is identity. We assume for simplicity that defendant would inevitably leave such a print, so that in this instance $P(H|G) = 1$.[22] If the trace left by the accused could vary in its characteristics, $P(H|G)$ would be less than 1.[23] It is also assumed that we know $P(H|NG)$, the probability that a palm print left by someone other than defendant would have the observed characteristics. Our problem is to express $P(G|H)$ in terms of $P(H|G)$ and $P(H|NG)$. That is, we want to know the probability that defendant used the knife, taking into account the chances that he or someone else left the palm print.

The probability of event G conditional on the occurrence of event H is, by definition in probability theory, the probability of the joint occurrence of G and H divided by the probability of H. In symbols:

$$P(G|H) = \frac{P(G \text{ and } H)}{P(H)}$$

This formula is intuitively reasonable because the probability

[21] T. STARKIE, A PRACTICAL TREATISE OF THE LAW OF EVIDENCE 751 n.h (9th Am. ed. 1869).

[22] Both $P(H|G)$ and $P(H|NG)$ are the probabilities that a print would have the observed characteristics, assuming that a right hand palm print was left by the person who used the knife. It is thus assumed that the leaving of a print is not per se evidence either for or against the defendant.

[23] See pp. 509–11 *infra*.

of G conditional upon H may be interpreted as the frequency with which G occurs out of all cases in which H occurs.[24] Applying the same definition:

$$P(H|G) = \frac{P(G \text{ and } H)}{P(G)}$$

so that P(G and H) can be written as P(H|G)P(G). In words, the probability of the joint occurrence of two events equals the probability of the first event times the probability of the second, conditional upon the occurrence of the first.[25]

P(H), the denominator of the fraction on the right hand side of the first equation, is the probability of finding the print. Since there is either identity or not — and since these alternatives are exhaustive — the sum of the chances of finding the print, given identity, and finding the print, given no identity, is the total probability of finding the print: [26]

$$P(H) = P(H \text{ and } G) + P(H \text{ and } NG)$$

Applying the definitions above for the joint occurrence of identity and finding the print, and of no identity and finding the print:

$$P(H) = P(G)P(H|G) + P(NG)P(H|NG)$$

Substituting these results for the numerator and denominator in the expression for P(G|H) yields Bayes' theorem:

$$P(G|H) = \frac{P(G)P(H|G)}{P(G)P(H|G) + P(NG)P(H|NG)}$$

This theorem is the desired result because it expresses P(G|H) in terms of P(H|G), P(H|NG), and P(G).[27] A way of looking at Bayes' theorem is to say that we start with some idea

[24] The numerator of the fraction given on the right hand side above is the probability of the joint occurrence of G and H; dividing by the denominator ensures that the total probability for all the cases in which H occurs will equal unity.

[25] In the special case when G and H are independent, P(H|G) = P(H) and P(G|H) = P(G). The probability of neither event is affected by the occurrence of the other. The foregoing then reduces to the "product" rule used in *Collins*: P(G and H) = P(G)P(H). The whole point of our case, of course, is that G and H are not independent.

[26] This follows from the "sum rule" which states that the probability of the occurrence of either of two mutually exclusive events (in this case use and non-use of the knife) is equal to the sum of the probabilities of those events.

[27] Derivations of Bayes' theorem may be found in elementary texts on probability theory, for example, J. FREUND, MATHEMATICAL STATISTICS 52–58 (1962). An extensive discussion of Bayes' theorem appears in Edwards, Lindman & Savage, *Bayesian Statistical Inference for Psychological Research*, 70 PSYCHOLOGICAL REVIEW 193 (1963). *See also* I. GOOD, PROBABILITY AND THE WEIGHING OF EVIDENCE ch. 6 (1950).

of the probability that defendant used the knife, P(G), and that our views are modified or weighted by the two probabilities associated with the print, P(H|G) and P(H|NG). Our final estimate of the chance defendant used the knife is our initial or "prior" view as modified by the statistical evidence.[28] It should be observed that P(G|H) does not depend on the size of the suspect population except as that factor may influence the prior probability or the frequency of the print.

The following table shows the resulting value of P(G|H) for various prior probabilities and statistical evidence. We assume that P(H|G) = 1. That is, any print left by the defendant on the knife would with certainty have the characteristics observed. The probability P(H|NG) is the frequency of the print in the suspect population.

TABLE I

POSTERIOR PROBABILITY P(G|H)

Frequency of Characteristics P(H|NG)	*Prior Probability P(G)*				
	.01	.1	.25	.50	.75
.50	.019	.181	.400	.666	.857
.25	.038	.307	.571	.800	.923
.1	.091	.526	.769	.909	.967
.01	.502	.917	.970	.990	.996
.001	.909	.991	.997	.9990	.9996

The table shows that even relatively high frequencies such as one in a hundred can lead to significant posterior probabilities if the prior view is at least one-fourth. For example, if the prior probability P(G) is .25 and the frequency of the observed characteristic in the population P(H|NG) is .01, then the posterior probability P(G|H) is .970. This is significant because evidence apart from statistics frequently will justify a fairly high prior probability of guilt. More modest probability estimates could thus have been used to make telling, even decisive, cases.

For example, in *People v. Risley*,[29] the issue was whether defendant had altered a court document by typing in the words "the same." Defendant was a lawyer and the alteration helped

[28] One may ask whether other "weightings" of prior probabilities and population frequency statistics would be justifiable. The answer is "no." By way of illustration, if the prior probability also represented the frequency of an event, Bayes' theorem would be the only correct way to reflect the joint effect of the two items of statistical information. And it has been shown that a subjective prior should be weighted no differently. *See* H. RAIFFA, DECISION ANALYSIS 124–27 (1968).

[29] 214 N.Y. 75, 108 N.E. 200 (1915).

his case. There was evidence tending to show that he had come to the clerk's office to examine the file (including the altered paper), then returned the next day and reexamined it. The state alleged that defendant had removed and replaced the document at these visits. This was physically possible.

Eleven defects in the typewritten letters on the court document were similar to those produced by defendant's machine. The prosecution called a professor of mathematics to testify to the chances of a random typewriter producing the defects found in the added words. The witness multiplied these component probabilities together to conclude that the joint probability of all defects was one in four billion. Given the magnitude of this estimate, the court was clearly correct, when it reversed, in objecting that the testimony was "not based upon observed data, but was simply speculative, and an attempt to make inferences deduced from a general theory in no way connected with the matter under consideration supply the usual method of proof." [30]

If the expert had adopted a Bayesian approach, he could have made good use of a justifiable probability estimate. On the evidence in *Risley* — excluding the evidence of similarity of defects — one might judge that there was at least a twenty-five percent chance that the alteration was typed on defendant's machine. Adding the information as to the defects and assuming that such defects would occur in fewer than one machine in a thousand, Bayesian analysis indicates a very high probability that defendant's machine was used. This is significant because an upper-bound estimate of one-in-a-thousand could probably have been supported — perhaps even on the basis of direct experience of the experts.[31]

IV.

Bayes' theorem demonstrates that even evidentiary traces linking a defendant to a crime which occur quite frequently can

[30] *Id.* at 85, 108 N.E. at 203. Apart from the problems arising from a blind use of the product rule, which have already been discussed, the testimony contained a defect of a rather general character. The expert testified that the probability estimate which he computed at one in four billion was "the probability of these defects being reproduced by the work of a typewriting machine, other than the machine of defendant . . . ." *Id.* When we remember that this number represents an estimate of the frequency in the population of typewriters with the specified defects, it is clear that the statement is incorrect. Assuming the expert's figure was right, the probability of duplication depends on two additional factors (which were not discussed): (1) the number of typewriters in the suspect population, and (2) the sharpness of the estimate.

[31] One of the experts called by the prosecution testified that he had examined 20,000 machines. 214 N.Y. at 83–84, 108 N.E. at 202.

help sustain an identification provided there is sufficient other evidence to connect the accused with the crime. This is a modest use which merely eliminates an unwarranted distinction between the force of statistical and other types of identification evidence. A stronger, more explicit use of the theorem is also possible. An expert witness could explain to jurors that their view of the statistical evidence should depend on their view of the other evidence. He might then suggest a range of hypothetical prior probabilities, specifying the posterior probability associated with each prior. Each juror could then pick the prior estimate that most closely matched his own view of the evidence. In *Risley*, the expert might have testified, for example, that if the jurors believed there was a fifty percent chance that the added words were typed on defendant's machine apart from the statistical evidence, they should believe that those chances were 999/1000 if they accepted the statistical evidence. To minimize the possibility that a prosecutor would prejudice the defendant's case by choosing only highly incriminating "hypothetical" prior probabilities, an expert so testifying should be required to show the posterior probabilities associated with a broad range of prior estimates. Such a procedure would also foreclose the chance that jurors would consider the expert as interjecting his own opinion as to the appropriate prior.[32]

Is there a need for this kind of explicit use? Arguably, there is. The statement that prints with particular characteristics occur with a frequency of one in a thousand persons means only that a defendant who has such a print is a thousand times more likely to have left it than someone selected at random from the population. By itself, this is not a meaningful statistic for measuring probability of guilt. As we have seen, a defendant could be a thousand times more likely to be guilty than someone selected at random and still more likely to be innocent than guilty. The comparison with a random selection is irrelevant. The jury's function is not to compare a defendant with a person selected randomly but to weigh the probability of defendant's guilt against the probability that *anyone* else was responsible. Bayes' theorem translates the one-in-a-thousand statistic into a probability statement which describes the probative force of that statistic.[33]

---

[32] Also, if hypothetical specification of probabilities of guilt was believed undesirable, the expert might testify that if — and only if — the jurors thought there was a substantial probability that the words were typed on defendant's machine without the statistics, they should think, if they believed the statistics, that it was very probable — in the neighborhood of 999 chances out of 1,000 — that the words were typed on defendant's machine.

[33] To test the utility of the explicit use of Bayes' theorem, the authors conducted an informal survey of intuition by using the facts in the case of the murdered woman. See pp. 496–97 *supra*. The subjects (admittedly not a random sample

Earlier legal commentators tended to take the view that mathematical probability was simply inapplicable to legal evidence.[34] This view appears to have been replaced by a greater receptivity, at least in theory, to probabilistic ideas expressed in mathematical form.[35] The change may reflect the growing prestige of the exact sciences and the inescapable fact that probability concepts lie inchoate behind many evidentiary standards of admissibility and proof. The standards "beyond a reasonable doubt" and "more likely than not," for example, import probabilistic notions.

The basis for opposing use of numerical probability in trials has rarely been stated with any clarity. In *Risley*, however, the court made its objection precise. It distinguished the accepted judicial use of life expectancy tables on the ground that probability concepts apply only to future events and thus cannot be used in determining guilt: "The fact to be established in this case was not the probability of a future event, but whether an occurrence asserted by the People to have happened had actually taken place." [36]

This view reflects a lack of familiarity with statistical inference. The difference between the future and the past is not significant to mathematical probability. A probabilistic analysis of the selection of a lottery ticket does not change when the ticket is drawn, but only when the results are known. Probability concepts are in fact routinely applied by statisticians to express uncertainties in measuring facts concerning a population.[37] Insofar as the distinction between future and past events is con-

---

from the population) were first given the facts, excluding the palm print information, and asked to assess the probability of defendant's guilt. They were then given the palm print statistics and asked for a reassessment. In all cases the prior probability was thought to be substantial in the sense we have defined it. In almost all cases the addition of the palm print evidence was thought to raise the probability of guilt, but assessments of the weight of this evidence varied widely, and the subjects were uncertain how to treat the new information. In most cases the assessments were not as great as they would have been if the probabilities had been computed in accordance with Bayes' theorem.

[34] *E.g.*, W. WILLS, AN ESSAY ON THE PRINCIPLES OF CIRCUMSTANTIAL EVIDENCE ILLUSTRATED BY NUMEROUS CASES 21 (3d ed. 1857).

[35] *See, e.g.*, C. MCCORMICK, EVIDENCE § 171 (1954).

[36] 214 N.Y. 75, 86, 108 N.E. 200, 203 (1915).

[37] A statistician estimating the average height of a population based on a sample whose average height was 5′ 6″ might express his conclusions as follows: the average height of the population lies in the range of 5′ 2″ to 5′ 10″ with a 95 percent probability. This statement does not imply that average height is a future event, but it does make a kind of prediction about the distribution of average heights in other such samples. The prediction is that if repeated samples were taken from this population and intervals constructed in the same way around the average height of each, we would expect that 95 out of 100 of the intervals would include the actual average height of the population.

cerned, there is no inherent reason why uncertainties in determining guilt could not also be expressed quantitatively.

The court's objection, however, does recall a difficult problem in attributing meaning to probability statements of the type used in Bayes' theorem. The concept of probability is usually defined in terms of relative frequencies of events. When we say that the probability of tossing heads with a coin is one-half, we mean that over a run of tosses heads will tend to come up half the time. It seems wholly artificial to apply a similar concept to the probability, say, that defendant used the knife. To do so would involve the assumption that the same testimony was repeated many times with the probability that defendant used the knife represented by the relative frequency of use in such repeated cases.

For this reason, some probabilists have argued that statements such as "there is a fifty percent chance that defendant used the knife" imply only a degree of belief in the proposition asserted and can not be interpreted as expressing a frequency. These kinds of estimates are said to express a "subjective," "intuitive," or "personal" probability. They have been defined in terms of the odds that a rational person acting after reflection and consistently would regard as fair in betting on the proposition.[38]

Although subjective probabilities can be used on this basis, we suggest that in the legal context they are likely to be interpreted as expressing a frequency, just as "the chances of heads is one-half" expresses a frequency.

When we say that defendant is guilty beyond a reasonable doubt, we mean that the evidence has brought us to a state of belief such that if everyone were convicted when we had such a belief the decisions would rarely be wrong. The "beyond reasonable doubt" standard thus groups together cases which are similar not because their facts are similar but because the degree of belief in guilt has passed a certain mark. A judge makes a similar classification on a lower level of probability when he allows a case to go to the jury or permits a verdict to stand. Thus, although it will usually be artificial to imagine a repetition of similar cases, one can nonetheless interpret subjective probability of (*e.g.*) guilt as the relative frequency of guilt over cases judged to be similar by the degree of belief they engender. The statement that "there is a fifty percent chance that defendant is guilty" thus means that if a jury convicted whenever the evidence generated a similar degree of belief in guilt, the verdicts in this group of cases would tend to be right about half the time. If this interpretation is accepted, then both subjective probability and probability as classically defined reflect frequencies of events.

[38] Discussed in 1 D. LINDLEY, INTRODUCTION TO PROBABILITY AND STATISTICS 32–34 (1965).

The classical approach, however, reflects occurrences (as, numbers of "heads") in a number of events (tossed coins) with some physical similarity. Subjective probabilities reflect occurrences (as, guilt) among events (cases) deemed similar because they generate similar degrees of belief. We suggest that this interpretation in fact represents the intuitive content of a subjective estimate of the probability of guilt.

Subjective probability estimates of guilt may vary widely depending on the person making the estimate.[39] This fact has often been raised as an objection to their use in scientific pursuits.[40] Whatever the validity of this objection in science, it does not have the same force in law. Varying judgments that reflect differing life experiences are accepted as an inevitable and even desirable aspect of the jury system. Moreover, in practice, differences among jurors who use Bayesian analysis will depend more on whether or not they believe the evidence establishing a subjective probability of guilt than on differences in the strength of their suspicions. If this evidence is disbelieved, the probability of defendant's guilt will be no stronger than that implied by defendant's belonging to the group of persons who have the trait in question, the size of that group being determined by statistical evidence. On the other hand, if the evidence is believed, both the prior suspicion and the statistical evidence will usually be strong enough so that, as Table I demonstrates, variations in the posterior probabilities will be small relative to variations in the strength of the suspicion.

---

[39] In most applications of Bayesian technique, the information supplied by the prior estimate of probability tends to weaken the inference to be drawn from the statistics. This is a consequence of the fact that most prior distributions assign equal or relatively equal probabilities to the hypotheses being tested. A prior of this type is called "gentle" if the probabilities assigned are not very different, and "flat" if they are exactly equal. The exactly equal case is sometimes equated with "no advance knowledge." If a flat prior were used in the case we have discussed, the probability of defendant's guilt would be 1/N where N is the total suspect population, and the Bayesian approach would reduce to the probabilities computed in the *Collins* line of cases. The proposed application is thus unusual in that a prior sufficiently ungentle to strengthen the statistical inference is a necessary step in the argument.

In their study of the disputed authorship problem in *The Federalist*, Mosteller and Wallace used the difference in rates of use of context-free words in papers of known authorship to determine odds for the papers of disputed authorship. They employed a flat or gentle prior to weaken the statistics on the ground that a weakening was justified to allow for effect of selection of words which were apparently good as discriminators from a large pool. *See* F. MOSTELLER & D. WALLACE, INFERENCE AND DISPUTED AUTHORSHIP: THE FEDERALIST 61 (1964).

[40] For over two centuries debate has swirled about the validity of Bayesian analysis in scientific pursuits and the prior probabilities with which it begins. The issues are discussed in JOINT STATISTICS SEMINAR, THE FOUNDATIONS OF STATISTICAL INFERENCE (G. Bernard & D. Cox eds. 1962).

Under certain restricted conditions, useful prior probabilities can be estimated on the basis of objective population statistics without resort to subjective evaluations. H. Steinhaus was the first to recognize this possibility. He computed a prior probability of paternity based on a sample of Polish paternity cases from the early 1950's.[41]

Under Polish family law, an accused is presumed to be the father of a child once it is proved that he is "the man who had sexual intercourse with the child's mother in the period from the 300th to the 180th day before its birth." [42] If such intercourse is proved, the burden shifts to the defendant to prove his non-paternity. If the child has a blood type which is not shared by the mother, the type could only have come from the father. In such a case, the defendant will be tested. If he does not have the type in question he is exonerated; if he has the type, the probability of guilt is increased but the possibility of innocence obviously is not foreclosed. These facts have made blood type evidence admissible in American courts solely for the purpose of exonerating the accused.[43]

Steinhaus used a Bayesian approach to compute the probability that defendants who failed to exonerate themselves by the blood test were in fact guilty. The prior probability he computed was the probability that the accused was the father after intercourse had been established but before the serological test. The posterior probability was the probability of paternity after the test. The significant aspect of Steinhaus' procedure is that he was able to use population statistics to calculate an estimate of the proportion of guilty fathers among those who were designated for the serological test even though no individuals (except those who were subsequently exonerated by the test) could be identified as guilty or innocent. To make the theory of his procedure clear, we simplify it slightly.

Different blood types have differing frequencies in the population. Let the type in question be called "A" and have the frequency f; the frequency of those who do not have this type is $1-f$. Consider the group of accused fathers who take a serological test because the child has blood type "A" which was not shared by the mother. If the mother's accusations were always right, the serological test would show every member of this group to have type "A" blood (although the converse of course is not true). If the mothers' accusations were always wrong, the members of this group would be a random sample from the popula-

[41] THE ESTABLISHMENT OF PATERNITY, PRACE WROCLAWSKIEGO TOWARZYSTWA NAUKOWEGO, ser. A., No. 32, at 5 (1954).

[42] Quoted at *id.*

[43] 1 J. WIGMORE, EVIDENCE § 165a (3d ed. 1940, Supp. 1962).

tion and the expected frequency of those with other than type "A" blood would be $1-f$. The difference between the actual rate of "A" blood is this accused group, and the population rate can be used to measure the accuracy of the accusations as a group. The more "A" blood, the more correct the accusations.[44]

Using the results of some 1,515 Polish paternity cases in which serological tests had been made, Steinhaus concluded that the prior probability of a true accusation in these cases in 1950 was about seventy percent (with perhaps less than complete fairness, this factor has been called "the veracity measure of women"). It is possible that a similar procedure can test other legal presumptions.

Steinhaus' program for developing his ideas was subsequently carried out by J. Lukaszewicz.[45] Using a different group of paternity cases in which serological tests had been made Lukaszewicz derived a frequency table by calculating the posterior probability for each case and counting the number of cases at each level of posterior probability. He showed that of a thousand cases there were 119 with zero posterior probability of paternity (exclusions); five cases with probability of paternity between 0.300 and 0.349, etc. In the highest probability category there were forty-five cases with probability of paternity between 0.950 and 1.000.

Armed with this information, Lukaszewicz then demonstrated the consequences of several different judicial strategies in a more specific way than has usually been thought possible. A judge could decide to dismiss all accusations where the posterior probability of paternity is less than one-half and to sustain all where the posterior probability is more than one-half. If a judge were to use such a fifty percent rule, the expected proportion of erroneous dismissals will equal the expected proportion of actual fathers for whom the probability of guilt is nevertheless calculated to be less than one-half.

Lukaszewicz calculated that the expected rate of erroneous dismissals would be 1.4 percent.[46] Similarly, the expected pro-

[44] Let p be the proportion of the accused group who are the fathers. Then $1-p$ is the proportion of innocents and $(1-p)(1-f)$ is the expected proportion of those accused who will be exonerated by the test. The ratio of the expected proportion of the accused group who will be exonerated to the proportion of those in the general population who do not have the blood type in question is $(1-p)(1-f)$. This ratio, however, is simply $1-p$, the prior probability of a false accusation. The key fact is that both numerator and denominator of the foregoing ratio can be estimated from objective population statistics.

[45] *O Dochodzeniu Ojcostwa* (*On Proving Paternity*), 2 ZASTOSOWANIA MATEMATYKI 349 (1955).

[46] Using the frequency table described above, Lukaszewicz calculated the sum of the expected number of actual fathers in each probability category from 0.0 to

portion of erroneous attributions of paternity would be 16.5 percent. Total error is thus 17.9 percent: the judge would be right about eighty-two percent of the time.

The choice of fifty percent probability as a criterion may seem to be an overly literal application of the preponderance of evidence rule. However, it can easily be shown that this rule of decision would result in less error than under any other posterior probability criterion. In addition, most of the error (ninety-two percent) would be erroneous attributions of paternity. These consequences might commend the fifty percent rule as a matter of social policy, although by using this rule the courts would consciously make more attributions of paternity than there were actual fathers among the accused.

A second possible strategy is intuitively appealing but on analysis is less acceptable. A judge who accepted the finding that seventy percent of the accused were guilty (the prior probability) might select the minimum probability for sustaining accusations that would result in a seventy percent conviction rate. On this condition the minimum probability required to sustain an accusation exceeds fifty percent. It is a characteristic of this method of decision that the proportion of erroneous dismissals equals the proportion of erroneous attributions. Calculations based on the Polish statistics show that both would be 10.9 percent. Total error is 21.8 percent so decisions would be right about seventy-eight percent of the time.

The total error resulting from such a decision rule is about four percentage points larger than under the fifty percent decision rule. Perhaps of greater significance, under this strategy the burden of mistakes would shift from the putative fathers and be equally divided between the parents. The proportion of erroneous dismissals would rise from 1.4 percent to 10.9 percent while the proportion of erroneous attributions would decline from 16.5 percent to 10.9 percent. This is probably a socially less desirable distribution of errors.

Lukaszewicz computed that the overall expected rate of error in court verdicts was between twelve and twenty-two percent. He did this by applying Steinhaus' method to determine the expected proportion of guilty defendants out of a group of some eight hundred, and comparing that proportion with the courts' conviction rate. The estimated rate of erroneous findings of paternity was between twelve and seventeen percent; the rate for wrongful dismissals was an estimated zero to five percent. The

0.5. The expected number of actual fathers in each category is equal to the product of the number of cases in that category times the probability of paternity for that category.

use of a fifty percent decision rule would thus probably not reduce the rate of wrongful attributions (16.5 percent as against twelve to seventeen percent) but might reduce the rate of wrongful dismissals (1.4 percent as against zero to five percent).

A judge might improve his performance further by using the statistically determined prior probability solely as a benchmark for what he believed to be an "average" case. Where in his opinion the facts showed that the case was either stronger or weaker than usual, he could subjectively adjust the prior accordingly. If his overriding policy was to minimize total error, the judge would continue to use the fifty percent rule, for this rule minimizes total error regardless of the prior probability value. If policy also encompassed weighting error for the mothers' benefit, the threshold value would have to be reduced from fifty percent when the prior was less than seventy percent; this adjustment would increase total error.

A Bayesian approach — particularly one that began with a subjectively estimated prior for the non-average case — would thus probably improve the performance of Polish judges in paternity cases. But whether or not there would be an improvement in judicial performance, the very high posterior probabilities of paternity demonstrated by the Polish statistics throw into question the rule in American jurisdictions that blood type evidence in such cases can be used only to absolve a defendant.

V.

In Table I, it was assumed that any right hand palm print left by the defendant on the knife would have the characteristics of the print actually observed. This was expressed as $P(H|G) = 1$. The assumption was made for simplicity and was probably reasonable as applied to fingerprints. But many other traces helpful in identification will vary because of variation within the suspected source. Thus a defendant may have only a certain chance of leaving a hair similar to one found at the scene of a crime. There may also be variations in reporting or measurement. In *Risley*, there were differences observed between the letters of the incriminating words and those subsequently produced by defendant's machine.[47] These differences were not sufficient to rule out defendant's machine as a source, but created some doubt, a diminished probability, that his machine produced the words. Similarly, in *Collins*, there were differences between the appearance of the defendants and the appearance of the guilty couple as described by the witnesses. The statistician ignored these dif-

[47] People v. Risley, 214 N.Y. 75, 85, 108 N.E. 200, 203 (1915).

ferences although they diminish the probability that the defendants would have been so described by the witnesses ($P(H|G)$).

Differences of the type described in *Collins* serve principally to cast some doubt on the conclusion of similarity. Usually there will be no hard data about the significance of the doubts raised, and $P(H|G)$ will have to be a guess. This might be embarrassing, except that Bayes' theorem indicates that even a substantial doubt is often insignificant to the result. Even if $P(H|G) = \frac{1}{2}$, instead of 1, a large reduction for this type of uncertainty, the posterior probabilities associated with the one-in-a-thousand statistic would deflate by only 1/1000.

But where there is significant variation within the suspected source, the doubt may well be so great as materially to decrease the likelihood of defendant's guilt. In these situations, some hard data about the variation must be used. Studies have shown, for example, that source variation is significant for hair but probably not for glass.[48] Investigators analyzing the source of hair have compared ten elements in the hair with the concentration of these elements in the hairs of a source's head. They have recommended the procedure that if the composition of the incriminating hair deviates from the average composition of a sample of defendant's hair so that such deviations would probably occur in, say, less than one percent of the defendant's hairs, then the incriminating hair is assumed not to be defendant's.[49] If the defendant's hair is "similar" to the incriminating hair (*i.e.*, the deviation is less than the selected standard) the probability that someone else left such a hair is computed by estimating the proportion of the suspect population whose hair was "similar" (by the same standard) to the incriminating hair.[50]

One difficulty with this procedure lies in its two-step approach. By itself, a decision that defendant's hair is "similar" to the incriminating hair, by the artificial standard selected, is without

[48] R. COLEMAN, F. CRIPPS, A. STIMSON, & H. SCOTT, *supra* note 3; R. COLEMAN & G. WOOD, THE VALUE OF TRACE ANALYSIS IN THE COMPARISON OF GLASS FRAGMENTS — A PRELIMINARY STUDY (U.K. Atomic Energy Auth., Atomic Weapons Research Establishment Report No. 03/68, 1968). The extent of source variation itself varies with the person and the element being considered. It sometimes approaches two-thirds of the variation of the element over the population. *See* R. COLEMAN, F. CRIPPS, A. STIMSON, & H. SCOTT, *supra*, tables 2 & 3, at 17, 18.

[49] One form of statistic used as a measure of difference is computed by taking the sum of the squared differences between the suspect's average measurements and the crime scene measurements divided by the standard deviations of these differences. For a discussion of various indices of this type see Parker, *The Mathematical Evaluation of Numerical Evidence*, 7 J. FORENSIC SCI. SOC. 134 (1967); Parker & Holford, *Optimum Test Statistics with Particular Reference to a Forensic Science Problem*, 17 APPLIED STATISTICS 237 (1968).

[50] For a discussion of the foregoing procedure, see Parker, *A Statistical Treatment of Identification Problems*, 6 J. FORENSIC SCI. SOC. 33 (1966).

probative significance. The admission into evidence of such a finding may be fatally prejudicial unless it is also shown that similar hairs are not common in the population. Nor can it be said that a finding of dissimilarity should exculpate an accused. A hair which may be quite rare for the accused, and in this sense unlikely to have come from him, may be still more unlikely to have come from someone else. Yet if a preliminary test of similarity has been adopted, the hypothesis that the hair is his would be rejected. By combining P(H|G) and P(H|NG) into a single formula, Bayes' theorem takes both factors into simultaneous account.

VI.

Determining P(H|NG) will usually require that inferences be drawn from samples taken from the general population. Complexities arise because characteristics useful for identification must be sufficiently rare so that they would appear either not at all or very infrequently (usually too infrequently for reliable statistical inference) in a sample of reasonable size. Thus, the expert in *Risley* might have testified, if asked, that of the thousands of Underwoods he had inspected, he had never seen one with the combination of defects in defendant's typewriter. We need a procedure for estimating the frequency of so rare an event.

Assume that out of a random sample of size n there are no occurrences of the trace in question. What inferences may be drawn from this fact? Since, in a criminal case, a defendant potentially identified by such a trace would not be prejudiced by too generous an estimate of its population frequency but only by one that was too small, we may compute and use an "upper-bound" estimate of the true frequency: one large enough so that there is only a negligible probability that the true frequency is larger. For example, if no identifying traces were found in a sample of one thousand, we could assume without prejudice to the accused that the frequency was ten percent, because the chances of the true frequency being larger than this would be negligible. It is possible to make this notion precise.[51] The fol-

[51] Let p denote the population frequency of the traces, and $q = 1 - p$ the frequency with which the traces do not appear. The probability of selecting a random sample (with replacement) of n elements none of which has the identifying trace is dependent upon q. We denote this probability as $P(n|q)$. For a given frequency, using the product rule:

$$P(n|q) = q^n$$

We seek $P(q|n)$, that is, the probability of q having a certain value given that n selections are made without finding a trace. In this way of looking at the problem, q is a random variable and n is a constant. To obtain $P(q|n)$ from $P(n|q)$ requires Bayes' theorem, and a prior probability $P(q)$. Using a form of the theorem

lowing table shows approximate upper-bound estimates for the population frequency p for varying sample sizes, based on the condition that there is only one chance in a hundred that p would be larger.

TABLE II

Upper-bound Estimates for $\bar{p}$

Sample Size	100	200	500	1,000	2,000
$\bar{p}$	.05	.02	.01	.005	.002

The table shows, for example, that finding no trace in a sample

---

generalized from that previously derived (p. 499) where we assume that q takes on a sequence of values between 0 and 1, we have:

$$(1) \qquad P(q|n) = \frac{P(q)P(n|q)}{\sum_{q=0}^{q=1} P(q)P(n|q)}$$

The "prior" probability here is $P(q)$. For a given value of q, $P(q)$ is the probability that q would have this value without considering the sample results. What values shall we assign to $P(q)$? In almost all cases, larger values of q would presumably have greater probability than small values since we are dealing with what are believed to be rare traces. Thus choice of a flat prior (*i.e.*, all possible values of q being deemed equally likely) is conservative in the sense that smaller values of q will be deemed more probable than they would be if a more realistic prior distribution had been used. Since an accused is favored by an estimate of q which reflects a greater probability of smaller values for q, the assumption of a flat prior should not be controversial.

We seek a value $\bar{q}$ to use as an estimate for q so that the sum of the posterior probabilities $P(q|n)$ for all values of q between o and $\bar{q}$ is less than a critical value, (*e.g.*, .01 or .05) which we denote as x. If $\bar{q}$ meets this condition, there is only x probability that the true value of q would be less than $\bar{q}$. Using the flat prior, the value of $\bar{q}$ satisfying this condition is given by the simple expression:

$$\bar{q} = x^{1/(n+1)}$$

The mathematical derivation of this result is as follows. Assuming a flat prior and recognizing that q is a continuous variate, Bayes' theorem becomes:

$$P(q|n) = \frac{P(n|q)}{\int_0^1 P(n|q)dq}$$

Substituting $P(n|q) = q^n$ and evaluating the intergal, we have:

$$P(q|n) = \frac{q^n}{\int_0^1 q^n dq} = (n+1)\ q^n$$

Then:

$$x = \int_0^{\bar{q}} P(q|n)dq = \int_0^{\bar{q}} (n+1)\ q^n dq = \bar{q}^{(n+1)}$$

Or:

$$\bar{q} = x^{1/(n+1)}$$

If x is given the commonly-used value .01, then, as we have determined it, $\bar{q}$

of one thousand justifies an assumption that the frequency of the trace is no larger than approximately five in one thousand. To justify assuming a frequency of one in a thousand (a statistic we have used in this article) it would be necessary to take a sample of about 4,650 without finding any trace.[52]

Cases where such large samples would be feasible probably are not common. *State v. Sneed*,[53] however, was such a case. There was evidence that the accused on occasion had used the name "Robert Crosset" and that on the day of the murder someone with the same name purchased a hand gun which, apparently, was the murder weapon. Were there two Robert Crossets? An expert witness examined telephone books in the area of the crime and found no "Crosset" in approximately 1,290,000 listings. He guessed that the frequency of "Crosset" must be about one in a million and estimated that the frequency of "Robert" was one in thirty. Using the product rule, he concluded that the frequency of "Robert Crosset" would be one in thirty million. In reversing the defendant's conviction, the Supreme Court of New Mexico did not object to the product rule, but did object to the use of "a positive number . . . on the basis of the telephone books when the name Robert Crosset was not listed in those books." [54]

The expert's conclusion was not justifiable. But by using the approach adopted here, he could have treated the telephone books as a large sample of the population which arguably was not biased with respect to the frequency of "Crosset" in at least the general area covered by the telephone books, and estimated this frequency at less than four in a million.[55] In civil cases, where it is desirable to balance the direction of errors in estimation, it would be more

---

is the smallest value for q (and hence $(1 - \bar{q})$ is the largest value for p) such that there is only one chance in a hundred that q would be smaller (or p larger). Where, as here, a binomial probability distribution is involved, a beta distribution is sometimes used for the prior probability. In this form $P(q)$ is proportional to $q^s (1 - q)^t$ where s and t are non-negative real numbers. *See* R. PRATT, H. RAIFFA, & R. SCHLAIFER, INTRODUCTION TO STATISTICAL DECISION THEORY ch. 11 (1965). If $P(q)$ is proportional to the beta distribution q^s, the estimate would be $q = x\,1/(n + s + 1)$ which is somewhat larger than the estimate given in the text. This illustrates the point previously made that the choice of a flat prior is conservative because it results in a smaller estimate for q and thus a larger estimate for p than if some other, more realistic choice, were made. Overestimation of p also results when the formulas here derived are applied to samples (taken without replacement) which are large relative to the population.

[52] With some increase in mathematical complexity, the method described here can be extended to the case where some elements of the random sample are found to have the identifying trace.

[53] 76 N.M. 349, 414 P.2d 858 (1966).

[54] *Id.* at 353, 414 P.2d at 861.

[55] $\bar{p} = 1 - \bar{q} = 1 - (.01)^{1/1,290,000} = .00000357$. Since "sampling" by the directories is without replacement, this result overestimates $\bar{p}$.

appropriate to use the expected value of the frequency of the trace rather than its upper-bound value. The difference between these two methods of estimation is illustrated by the fact that the expected frequency of "Crosset" would be less than one in a million, or about four times smaller than the upper-bound value.[56]

VII.

Where the incriminating trace consists of a number of elements which individually appear with some frequency in the sample, the information provided by these frequencies can be combined to generate even more powerful results than can be inferred from the nonappearance in the sample of the trace as a whole. In *Collins* the probabilities of the individual elements were simply multiplied together. As we have seen, the validity of this method depends on the assumption of independence. In most cases independence cannot be assumed. One must use a different technique, one which makes allowance for possible correlations among elements.

The product rule leads to a probability estimate for a compound event which is consistent with the probabilities of the elements comprising the event in the sense that the total probability for all mutually exclusive compound events in which the individual element occurs equals the probability of the individual element. For example, if we throw three dice, our estimate of the probability of three sixes should be consistent with our estimate of the probability of a six with each of the dice. This means that the sum of the probabilities for all combinations which include a six on the first die should be equal to the probability of six on that die. Similarly for each of the other dice. If the dice were thrown one hundred times, we might not see enough three sixes to be able to estimate directly the frequency of this event (other than by the upper-bound procedure already discussed), but we would see a sufficient number of sixes for each die to be able to estimate their frequencies with some confidence. We would use the product rule to obtain an estimate for the frequency of three sixes, the compound event, which was consistent in the sense described with the frequencies of the individual events, a six on each die.[57]

---

[56] The expected value of q would be $\int_0^1 (n+1)q^{n+1}dq = \frac{n+1}{n+2}$ and consequently the expected value of $p = 1/n+2$.

[57] The difference between the upper-bound method of estimation which makes use solely of the nonappearance of the trace, and an estimate based on the frequencies of elements of the trace may be illustrated with this dice example. If we throw three dice one hundred times without observing three sixes (a probable occur-

Similarly, if there are correlations between pairs of individual elements, we cannot use the product rule unmodified, but we can look at the frequencies of all possible pairs of elements and make our estimate consistent with the frequencies of pairs just as the product rule does with the frequencies of individual elements when there are no correlations. Estimates made in this way require a multiple iteration technique by which the solution appears as the end product of a series of successive approximations instead of a simple multiplication together of individual probabilities.[58] In principle, however, the method is the same. For example, if in *Risley* we anticipated that defects in individual letters were correlated as pairs we would look at the frequency of each such pair and estimate the probability of the occurrence of all defects in a way which was consistent with the frequency of the individual pairs.

The same method can be extended to higher order correlations and, for example, estimates made on the basis of the frequency of triplets. But the data reflecting the frequency of complex events thin out rapidly and we soon find too few cases or perhaps none with the requisite combinations of individual elements. The fewer the number of events the weaker the precision of the estimate. Thus, the problem of independence of factors which the court rightly criticized in *Collins* may be pushed back but not altogether eliminated. If estimates are sought to be based on the frequencies of elements of a trace, the assumption must be used that at least some higher order correlations do not exist. This assumption will appear more or less reasonable depending upon the circumstances. It might be fairly strong in *Collins* where the most significant effects might be correlations of pairs of attributes (*i.e.*, beard and mustache) but much weaker in *Risley* where the defects might be linked to the age of the typewriter.

---

rence), our upper-bound estimate for the frequency of three sixes would be approximately 5/100 (see Table II). If the dice were true, our estimate using the product rule would be in the neighborhood of 1/216, which is approximately ten times smaller than the upper-bound estimate.

[58] *See* Y. BISHOP, MULTIDIMENSIONAL CONTINGENCY TABLES: CELL ESTIMATES (Ph.D. Thesis, Dep't of Statistics, Harvard University 1967); Mosteller, *Association and Estimation in Contingency Tables*, 63 J. AM. STAT. ASS'N 1, 18–27 (1968). A related approach which makes use of Bayesian techniques appears in Fienberg & Holland, *Methods for Eliminating Zero Counts in Contingency Tables*, THE BIOMETRIC SOCIETY SYMPOSIUM ON RANDOM COUNTS IN SCIENTIFIC WORK (to be published). The method outlined here has been applied in a major study of the possibility of a causal relationship between halothane anesthesia and massive hepatic necrosis following surgery. *See Summary of the National Halothane Study*, 197 J. AM. MED. ASS'N 775 (1966). Forensic science applications would appear to offer a good occasion for experimenting further with this technique.

VIII.

In the *Howland Will* case,[59] a mathematician, Professor Benjamin Peirce of Harvard, applied the product rule to strokes of authentic and disputed signatures and concluded that their similarities were a phenomenon which could occur only once in the number of times expressed by the thirtieth power of five. "This number," he testified, "far transcends human experience. So vast an improbability is practically an impossibility. Such evanescent shadows of probability cannot belong to actual life. They are unimaginably less than those least things the law cares not for." [60]

Numbers of this magnitude have been a consistent feature of cases like *Collins*. Unsupported, and essentially unsupportable by data, they are likely to remain theoretical abstractions signifying little more than the expert's judgment that the event was unique. But the intrusion of such "evanescent shadows" intimidates and stultifies thought, and may generate skepticism in the more sophisticated. In cases like *Collins* expert judgment will rarely be improved or better communicated by statistics, and may be distorted.

But when the event is not expected to be unique, so that the expert should say only that it is to some degree rare, there is significant value in a statistical rendering of his opinion. Rarity will mean different things to different people. Without further explanation, a juror has no way of assessing the significance of the evidence. He might bring his own experience to bear when commonplace traits are involved, but he will be baffled by the technical data that are likely to become increasingly involved in future cases. If true judgment is to be exercised, he must know something more precise about rarity than the word alone can communicate.

There has been uncertainty in the opinions dealing with statistics about the probative significance of events not unique in the population. We have argued that it is appropriate to translate frequencies of such events into a probability statement by combining them with prior probabilities through the use of Bayes' theorem. The results confirm our intuitive notion that a trait need not be unique in the population in order to have probative significance. One need not actually inject Bayes' theorem into the courtroom to make use of this result, for it justifies allowing statistical evidence introduced without explicit use of a prior to have its natural impact on the jury. We have advocated, however,

---

[59] Robinson v. Mandell, 20 Fed. Cas. 1027 (No. 11959) (C.C.D. Mass. 1868), *discussed in The Howland Will Case*, 4 AM. L. REV. 625 (1870).

[60] *The Howland Will Case*, 4 AM. L. REV. 625, 649 (1870).

making explicit use of Bayes' theorem in order to translate the data into a form congenial to scrutiny by jurors.

In rejecting unjustifiable statistics, the courts have expressed concern that, for various reasons, statistical methods might be unfair to defendants. But while abuse is of course possible, mathematics correctly used should lead to a fairer evaluation of identification evidence.

In determining the population frequency of an incriminating trace, the choice is between expertise expressed in the traditional nonquantitative way (*e.g.*, the trace is rare), and objective studies with results reported in quantitative terms (*e.g.*, one in a thousand). A defendant will generally be favored by the quantitative study because, as we have already observed, the conclusions are likely to be less firm than those of "pure" judgment. In addition, the expert's method will be more exposed to examination and attack. It seems clear that quantitative expression of trace frequencies would not be unfair.

Bayesian analysis adds a dimension to the problem. There is a danger that in quantifying their suspicions jurors will overstate their convictions and thus be led by the mathematics to conclude guilt to be more probable than if they had considered the same evidence without quantification. On the other hand, a juror forced to derive a quantitative measure of his suspicion on the basis of the evidence at trial is likely to consider that evidence more carefully and rationally, and to exclude impermissible elements such as appearance or popular prejudice. Moreover, Bayesian analysis would demonstrate that the evidentiary weight of an impressive figure like one in a thousand — which might otherwise exercise an undue influence — would depend on the other evidence in the case, and might well be relatively insignificant if the prior suspicion were sufficiently weak. Probably the greatest danger to a defendant from Bayesian methods is that jurors may be surprised at the strength of the inference of guilt flowing from the combination of their prior suspicions and the statistical evidence. But this, if the suspicions are correctly estimated, is no more than the evidence deserves.