Scientific Heat about Cold Hits

Keith Devlin[*]

UNFINISHED DRAFT
COMMENTS WELCOMED

January 15, 2007

Abstract

This paper presents a study, from a mathematician's perspective, of the current ongoing debate as to how to calculate the significance of a DNA profile match in a "Cold Hit" case, where the match is the result of a search through a DNA database, and what statistical information about the database identification may be presented in court as evidence. At present, such evidence may be (though often is not) excluded from court proceedings in jurisdictions that adhere to the 1923 *Frye* ruling that only scientific evidence may be admitted that uses methods on which the scientific community has reaches a consensus. Much of the legal debate has centered around the government's insistence that it present the RMP (random match probability) in cold hit cases, a position that we argue strongly against.

We use a particular current case as an illustrative example, but the focus of the paper is the general mathematical issues involved in such cases.

The paper is in part expository, written both to make the mathematical community more aware of this important legal issue, and to provide lawyers and others in the legal profession with the necessary background to understand the main statistical issues involved. We do however present a proposed resolution to the key disagreement between the two main protagonists in the statistical debate — the adherents of the procedures recommended in the National Research Council's 1992 and 1996 reports on DNA profiling evidence, and those who favor the Bayesian approach advocated by Balding and Donnelly.

United States of America versus Raymond Jenkins

On June 4 1999, police officers in Washington, D.C. found the body of Dennis Dolinger, aged 51, at his home in Capitol Hill. He had been stabbed multiple times — at least 25 according to reports — with a screwdriver that penetrated into the brain.

Dolinger had been a management analyst at the Washington Metropolitan Area Transit Authority. He had lived on Capitol Hill for 20 years and was active in the community. In particular, he was a neighborhood politician, having been elected

[*] Cordura Hall, Stanford University, Stanford, CA 94305-4115.

ANC 6B12 representative in 1993, and had taken a strong stand against drug dealing in the area. He had a wide network of friends and colleagues across the city. He was openly gay and lived with a housemate, whom police did not consider a suspect.

Police found a blood trail leading from the basement where Dolinger was discovered to the first- and second-floors of his house and to the front walkway and sidewalk. Bloody clothing was found in the basement and in a room on the second floor. Police believed the some of the bloodstains were those of the murderer, who was cut during the assault. Dolinger's wallet, containing cash and credit cards, had been taken, and his diamond ring and gold chain were missing.

The police quickly identified several suspects: Dolinger's former boyfriend, who had assaulted Dolinger in the past and had left the D.C. area around the time police discovered the body, a man who was observed fleeing from Dolinger's house but did not call the police, neighborhood drug dealers, including one in whose murder trial Dolinger was a government witness, neighbors who had committed acts of violence against Dolinger's pets, various homeless individuals who frequently visited Dolinger, and gay men whom Dolinger had met at bars through Internet dating services.

By far the strongest lead was when a man called Stephen Watson used one of Dolinger's credit cards at a hair salon and department store in Alexandria within fifteen hours of Dolinger's death. Watson was a drug addict and had a long criminal record including drug offenses, property offenses, and assaults. Police spoke with a witness who knew Watson personally and saw him on the day of the murder in the general vicinity of Dolinger's home, "appearing nervous and agitated" with "a cloth wrapped around his hand" wearing a "t-shirt with blood on it." Another witness also saw Watson in the general vicinity of Dolinger's home on the day of the murder, and noted that Watson had several credit cards with him.

On June 9, police executed a search warrant at Watson's house in Alexandria, where they found some personal papers belonging to Dolinger. They also noticed that Watson, who was present during the search, had a cut on his finger "that appeared to be several days old and was beginning to heal." At this point, the police arrested Watson. When questioned at the police station, Watson "initially denied knowing the decedent and using the credit card" but later claimed that "he found a wallet in a back pack by a bank along side a beige colored tarp and buckets on King Street" in Alexandria. Based on those facts, the police charged Watson with felony murder.

In the meantime, the FBI had extracted and analyzed DNA from various blood samples collected from the crime scene. The Agency's DNA laboratory determined the DNA profile of the samples. (Specifically, they determined the profile at the thirteen "CODIS loci" that they used to generate a "DNA profile" —

see presently.) When the profile from the blood samples failed to match that of Watson, the US Attorney's Office dropped the case against Watson, who was released from custody.

At that point, the FBI ran the crime scene DNA profile through its database of DNA profiles of known offenders (a database in the FBI's CODIS system — see momentarily) to see if a match could be found, but the search came out negative.

Six months later, in November 1999, the DNA profile of the unknown contributor of the blood evidence was sent to the Virginia Division of Forensic Science, where a computer search was carried out comparing the profile against the 101,905 offender profiles in the Virginia DNA databank. This time a match was found — albeit at only eight of the thirteen CODIS loci, since the Virginia database listed profiles based on those eight loci only.

The (eight loci) match was with a man listed as Robert P. Garrett. A search of law enforcement records revealed that Robert P. Garrett was an alias used by Raymond Anthony Jenkins, who was serving time in prison for second-degree burglary — a sentence imposed following his arrest in July 1999, a few weeks after Dolinger was murdered. From that point on, the police investigation focused only on Jenkins.

On November 18, 1999, police interviewed a witness — a man who was in police custody at the time with several cased pending against him — who claimed to know Jenkins. This witness reported that on the day after Dolinger's death he had seen Jenkins with several items of jewelry, including a ring with some diamonds and some gold chains, and more than $1,000 in cash. Jenkins also appeared to have numerous scratches or cuts to his face, according to government documents.

Seven days later the police executed a search warrant on Jenkins for blood samples. The samples were sent to the FBI's forensic science lab for comparison. In late December 1999, Jenkins' blood samples were analyzed and profiled on the FBI's thirteen CODIS loci, the eight used by the Virginia authorities plus five others. According to a police affidavit, the resulting profile was "positively identified as being the same DNA profile as that of the DNA profile of the unknown blood evidence that was recovered from the scene of the homicide." The FBI analysis identified Jenkins's blood on a pair of jeans found in the basement near Dolinger, a shirt found in the upstairs exercise room, a towel on the basement bathroom rack, the sink stopper in the sink of the same bathroom and a railing between the first and second floors of the residence. Based on frequencies estimated from a database of 210 self-reported African-Americans, the FBI estimated that the probability that a random person selected from the African-American population would share Jenkins' profile (the so-called "random match probability", of which more presently) is 1 in 26 quintillion. Based on that information, an arrest warrant was issued, and Jenkins was arrested on January 13, 2000.

At the time of writing this article, in late 2005, the case against Jenkins has yet to go to trial. The main issue of contention between the government prosecutors and the public defenders appointed to represent Jenkins is whether the DNA profile produced by the FBI serves to identify Jenkins as the person who murdered Dennis Dolinger, and (accordingly) whether the result of the DNA profiling can be admitted in court as evidence. The question is not one of biochemistry. All parties involved agree that the laboratory science is sound and reliable, and that Jenkins' DNA profile (on the 13 CODIS loci) matches that of samples taken from the crime scene just after the murder. What is in dispute is what that match signifies. And that turns out to be a question of mathematics.

Before we look at the mathematical issues, however, we need to familiarize ourselves with the technique of DNA profiling.

DNA profiling

The DNA molecule comprises two long strands, twisted around each other in the now familiar double-helix structure, joined together in a rope-ladder-fashion by chemical building blocks called bases. (The two strands constitute the "ropes" of the "ladder", the bonds between the bases its "rungs".) There are four different bases, adenine (A), thymine (T), guanine (G), and cytosine (C). The human genome is made of a sequence of roughly three billion of these base-pairs. Proceeding along the DNA molecule, the sequence of letters denoting the order of the bases (a portion might be … AATGGGCATTTTGAC …) provides a "readout" of the genetic code of the person (or other living entity). It is this "readout" that provides the basis for DNA profiling.

Using today's techniques (let alone those used at the time the Jenkins' investigation was carried out), it would be totally impractical to do a DNA comparison by determining all three billion letters. What is done instead is to examine a very small handful of sites of variation.

DNA is arranged into large structural bodies called chromosomes. Humans have 23 pairs of chromosomes which together make up the human genome. One chromosome in each pair is inherited from the mother and the other from the father. This means that an individual will have two complete sets of genetic material. A "gene" is really a location (locus) on a chromosome. Some genes may have different versions, which are referred to as "alleles." A pair of chromosomes have the same loci all the way along their length, but may have different alleles at some of the loci. Alleles are characterized by their slightly different base sequences and are distinguished by their different phenotypic effects. Some of the genes studied in forensic DNA tests have as many as 35 different alleles in the population.

Most people share very similar gene sequences, but some regions of DNA sequence vary from person to person with high frequency. Comparing variation in these regions allows scientists to answer the question of whether two different DNA samples come from the same person.

The profiling technique used by the FBI and other law enforcement authorities depends on the fact that the variability is manifested by differences in the length, measured by the number of bases or the number of times a given sequence repeats, between pre-specified locations. This procedure yields two measurements for each sample for each locus, one for the father's side and one for the mother's side. The length of DNA fragments can be measured precisely. In comparing two samples at a given locus, if the pair of measurements from one sample is the same as the pair of measurements from the other, the profiles are said to match at that locus; otherwise, they are said not to match at that locus. If the two profiles match at each of the loci examined, the profiles are said to match. If the profiles fail to match at one or more loci, then the profiles do not match, and it is virtually certain that the samples do not come from the same person.

A match does not mean that the two samples must absolutely have come from the same source; all that can be said is that, so far as the test was able to determine, the two profiles were identical, but it is possible for more than one person to have the same profile across several loci. At any given locus, the percentage of people having DNA fragments of a given length, in terms of base pairs, is small but not zero. DNA tests gain their power from the conjunction of matches at each of several loci; it is extremely rare for two samples taken from unrelated individuals to show such congruence over many loci.

The FBI's forensic DNA identification system CODIS examines thirteen such regions in the genome. Sequences in these special regions involve multiple repetitions of short combinations of letters, such as GATA. Easily detectable differences between people lie in the number of repeats that occur in both copies of their DNA in these regions. For example, at one of these regions a person might have inherited four repeats (GATAGATAGATAGATA) from their father and six repeats (GATAGATAGATAGATAGATAGATA) from their mother at the same location in the genome. Another person might inherit eight repeats (GATAGATAGATAGATAGATAGATAGATAGATA) from their father and five repeats (GATAGATAGATAGATAGATA) from their mother.

When two randomly chosen DNA samples match completely in a large number of regions, such as the 13 used in the FBI's system, the probability that they could have come from two unrelated people is virtually zero. This fact makes DNA identification extremely reliable (when performed correctly). The degree of reliability is generally measured by using probability theory to determine the likelihood of finding a particular profile among a random selection of the population.

For example, consider a profile based on just three sites. The probability that someone would match a random DNA sample at any one site is roughly one in ten (1/10).[1] So the probability that someone would match a random sample at three sites would be about one in a thousand:

$$1/10 \times 1/10 \times 1/10 = 1/1{,}000.$$

Applying the same probability calculation to all 13 sites used in the FBI's CODIS system would mean that the chances of matching a given DNA sample at random in the population are about one in ten trillion:

$$(1/10)^{13} = 1/10{,}000{,}000{,}000{,}000.$$

This figure is known as the *random match probability* (RMP). Since it is computed using the product rule for multiplying probabilities, it assumes that the patterns found in two distinct sites are independent. During the early days of DNA profiling, this was a matter of some considerable debate, but by and large that issue seems to have died away.

In practice, the actual probabilities vary, depending on several factors, but the figures calculated above are generally taken to be a fairly reliable indicator of the likelihood of a random match. That is, the RMP is generally taken as a good indicator of the rarity of a particular DNA profile in the population at large, although this interpretation needs to be viewed with care. (For example, identical twins share almost identical DNA profiles.)

The denominator in the FBI's claimed figure of 1 in 26 quintillion in the *Jenkins* case seems absurdly high. Although I have no solipsistic tendencies, I don't think I could claim with that kind of certainty that either you or Raymond Jenkins or the FBI exists outside my imagination and that I am not merely a brain in a vat. In fact, even the one in ten trillion figure given by the more widely accepted calculation is surely several orders of magnitude less than the likelihood of other errors, such as contamination errors during sample collection or laboratory errors during the analysis process.

Nevertheless, whatever actual numbers you compute, it is surely the case that a DNA profile match on all thirteen of the sites used by the FBI is a virtual certain

[1] Profile match probabilities are based on empirical studies of allele frequencies of large numbers of samples. The figure 1/10 used here is widely regarded as being a good representative figure. See for example http://www.koshlandscience.org/exhibitdna/crim03.jsp

identification — *provided that the match was arrived at by a process consistent with the randomness that underpins the RMP*. (And there's the rub. History is littered with examples of how statistics can mislead, and mislead badly, if the appropriate populations are not random. A famous case is the landslide victory of Harry S. Truman in the 1948 Presidential election, when all the opinion polls showed that his opponent, Thomas Dewey, had a strong lead. The polls were not conducted with a sufficiently random sample of electors.)

The CODIS system

In 1994, recognizing the growing importance of forensic DNA analysis, Congress enacted the DNA Identification Act, which authorized the creation of a national convicted offender DNA database and established the DNA Advisory Board (DAB) to advise the FBI on the issue. DAB members were appointed by the director of the FBI from a list of experts nominated by the National Academy of Sciences and professional forensic science societies.

CODIS, the FBI's DNA profiling system (the name stands for COmbined DNA Index System) had been started as a pilot program in 1990. The system blends computer and DNA technologies to provide a powerful tool for fighting crime. The CODIS DNA database is comprised of four categories of DNA records:

- Convicted Offenders - DNA identification records of persons convicted of crimes;
- Forensic - Analyses of DNA samples recovered from crime scenes;
- Unidentified Human Remains - Analyses of DNA samples recovered from unidentified human remains;
- Relatives of Missing Persons - Analyses of DNA samples voluntarily contributed from relatives of missing persons.

The CODIS database of convicted offenders currently contains over 2.7 million records.

The DNA profiles stored in CODIS are based on thirteen specific loci, selected because they exhibit considerable variation among the population.

CODIS utilizes computer software to automatically search these databases for matching DNA profiles.

CODIS also maintains a Population file, a database of anonymous DNA profiles used to determine the statistical significance of a match.

CODIS is not a comprehensive criminal database, but rather a system of pointers; the database only contains information necessary for making matches. Profiles stored in CODIS contain a specimen identifier, the sponsoring laboratory's identifier, the initials (or name) of DNA personnel associated with the analysis, and the actual DNA characteristics. CODIS does not store criminal history information, case-related information, social security numbers, or dates-of-birth.

Using DNA profiling

Suppose that, as often occurs, the authorities investigating a crime obtain evidence that points to a particular individual as the criminal, but fails to identify the suspect with sufficient certainty to obtain a conviction. If the suspect's DNA profile is in the CODIS database, or else a sample is taken and a profile prepared, it may be compared with a profile taken from a sample collected at the crime scene. If the two profiles agree on all thirteen loci, then for all practical — and all legal — purposes, the suspect can be assumed to have been identified with certainty. The random match probability (one in ten trillion) provides a reliable estimate of the likelihood that the two profiles came from different individuals. (The one caveat is that relatives should be eliminated. This is not always easy, even for close relatives such as siblings; brothers and sisters are sometimes separated at birth and may not even be aware that they have a sibling, and official records do not always correspond to reality.)

Of course, all that a DNA match does is identify — within a certain degree of confidence — an individual whose DNA profile was that same as that of a sample (or samples) found at the crime scene. In of itself, it does not imply that the individual committed the crime. Other evidence is required to do that. For example, if semen taken from the vagina of a woman who was raped and murdered provides a DNA profile match with a particular individual, then, within the calculated accuracy of the DNA matching procedure, it may be assumed that the individual had sex with the woman not long before her death. Other evidence would be required to conclude that the man raped the woman, and possibly further evidence still that he subsequently murdered her. A DNA match is just that: a match of two profiles.

As to the degree of confidence that can be vested in the identification of an individual by means of a DNA profile match obtained in the above manner, the issues to be considered are:

- The likelihood of errors in collecting (and labeling) the two samples and determining the associated DNA profiles

- The likelihood that the profile match is purely coincidental.[2]

[2] As will be explained later, care is required in interpreting this requirement in terms of exactly what numerical probability is to be computed.

A likelihood of one in ten trillion attached to the second of these two possibilities (such as is given by the RMP for a 13-loci match) would clearly imply that the former possibility is far more likely, since hardly any human procedure can claim a one in ten trillion fallibility rate. Put differently, if there is no reason to doubt the accuracy of the sample collections procedures and the laboratory analyses, the DNA profile identification could surely be viewed with considerable confidence.

There is still some doubt regarding the use of the RMP to obtain a reliable indicator of an accidental match, computed as it is on the basis of our current scientific understanding of genetics. The RMP calculation does, after all, require mathematical independence of the loci — an extremely demanding condition — in order to be able to apply the product rule. It should be noted that a recent analysis of the Arizona convicted offender data base (a database that uses the 13 CODIS loci) revealed that among the approximately 65,000 entries listed there were 144 individuals whose DNA profiles match at 9 loci (including one match between individuals of different races, one Caucasion, the other African American), another few who match at 10 loci, one pair that match at 11, and one pair that match at 12. The 11 and 12 loci matches were siblings, hence not random. But matches on 9 or 10 loci among a database as small as 65,000 entries cast considerable doubt on figures such as “one in ten trillion” for a match that extends to just 3 or 4 additional loci.[3] [4]

But a one-in-a-trillion likelihood is massive overkill. Absent any confounding factors, a figure of one in a million or one in ten million (say) would surely be enough to determine identity with virtual certainty. Hence, all of the above cautions notwithstanding, it seems reasonably clear that (relatives aside) a 13-loci match can be taken as definitive identification — provided that, *and this is absolutely critical to the calculation and use of the RMP*, the match is arrived at by comparing a profile from a sample from the crime scene with a profile taken from a sample from a suspect *who has already been identified by means other than his or her DNA profile*.

But this is not what happened in the Jenkins case. There, Jenkins became a suspect solely as a result of trawling through a DNA database (two databases, in fact) until a match was found — the so-called “Cold Hit” process.

Cold Hit searches

[3] The matches may be due to individuals who are fairly closely related. Family relationships are not always known to the individuals themselves, nor to the authorities, nor even ever recorded.

[4] The situation is more subtle than might first appear. When the mathematics is done with care, the Arizona results are not at variance with what the mathematics predicts. The problem is in how people interpret the math. I’ll come back to this issue later.

In general, a search through a DNA database, carried out to see if a profile can be found that matches the profile of a given sample — say a sample obtained from a crime scene — is called a Cold Hit search. A match that results from such a search would be a "*cold* hit" because, prior to the match the individual concerned was not a suspect.

For example, CODIS enables government crime laboratories at a state and local level to conduct national searches that might reveal, say, that semen deposited during an unsolved rape in Florida could have come from a known offender from Virginia.

As in the case where DNA profiling is used to provide identification of an individual who was already a suspect, the principal question that has to be (or at least *should* be) asked after a cold hit search has led to a match (a "hit") is: Does the match indicate that the profile in the database belongs to the same person whose sample formed the basis of the search, or is the match purely coincidental? At this point, the waters rapidly become very murky.

To illustrate the problems inherent in the Cold Hit procedure, consider the following analogy. A typical state lottery will have a probability of winning a major jackpot around 1 in 35,000,000. To any single individual, buying a ticket is clearly a waste of time. Those odds are effectively nil. But suppose that each week, at least 35,000,000 people actually do buy a ticket. (This is a realistic example.) Then every one to three weeks, on average, someone will win. The news reporters will go out and interview that lucky person. What is special about that person? Absolutely nothing. The *only* thing you can say about that individual is that he or she is the one who had the winning numbers. You can make absolutely *no* other conclusion. The 1 in 35,000,000 odds tell you *nothing* about any other feature of that person. The fact that there is a winner reflects the fact that 35,000,000 people bought a ticket — and *nothing* else.

Compare this to a reporter who hears about a person with a reputation of being unusually lucky, goes along with them as they buy their ticket, and sits alongside them as they watch the lottery result announced on TV. Lo and behold, that person wins. What would you conclude? Most likely, that there has been a swindle. With odds of 1 in 35,000,000, it's impossible to conclude anything else in this situation.

In the first case, the long odds tell you *nothing* about the winning person, other than that they won. In the second case, the long odds tell you *a lot*.

A Cold Hit measured by RMP is like the first case. All it tells you is that there is a DNA profile match. It does not, in of itself, tell you anything else, and certainly not that that person is guilty of the crime.

On the other hand, if an individual is identified as a crime suspect by means other than a DNA match, then a *subsequent* DNA match is like the second case. It tells you a *lot*. Indeed, assuming the initial identification had a rational, relevant basis (like a reputation for being lucky in the lottery case), the long RMP odds against a match could be taken as conclusive. But as with the lottery example, in order for the long odds to have (*any*) weight, the initial identification has to be *before* the DNA comparison is run (or at least demonstrably independent thereof). Do the DNA comparison first, and those impressive sounding long odds may be totally *meaningless*, simply reflecting the size of the relevant population, just as in the lottery case.

Not everyone agrees with the above analogy — at least, they do not agree with the conclusions regarding the inapplicability of the RMP in the case of a cold hit match. In particular, the prosecution in the Jenkins case have argued consistently that the RMP remains the only statistic that needs to be presented in court to provide a metric for the efficacy of a DNA match. The defense in that case have argued equally consistently that the RMP is so misleading, particularly for laypersons, that it should not be presented to a jury in court — and in that case so far it has not.

Of course, in a legal system based on adversarial process, such as ours (the USA), one would expect the two sides in a case like *Jenkins* to take such diametrically opposed positions, particularly given that the case is likely to establish a significant legal precedent. What makes the situation interesting from a mathematical point of view is that each side has presented testimony in its favor from some decidedly well qualified statisticians.

From a legal standpoint, the very existence of a scientific controversy of this nature may be sufficient to keep the RMP out of the court proceedings, at least in the District of Columbia and in any of the 17 states that follow a 1923 ruling by a federal D.C. Court — 293 F. 1013 (D.C. Cir. 1923) — known generally as the *Frye* "general acceptance" test. This says that admissible scientific evidence must be based on a "well-recognized scientific principle or discovery [that is] sufficiently established to have gained general acceptance in the particular field to which it belongs".[5] For a scientific theory or technique to be a basis for

[5] The *Frye* case considered the admissibility of evidence obtained using an early form of lie detector based on changes in systolic blood pressure. Counsel for defendant, arguing for admitting the evidence, stated in their brief to the court: "The rule is that the opinions of experts or skilled witnesses are admissible in evidence in those cases in which the matter of inquiry is such that inexperienced persons are unlikely to prove capable of forming a correct judgment upon it, for the reason that the subject-matter so far partakes of a science, art, or trade as to require a previous habit or experience or study in it, in order to acquire a knowledge of it. When the question involved does not lie within the range of common experience or common knowledge, but requires special experience or special knowledge, then the opinions of witnesses skilled in that particular science, art, or trade to which the question relates are admissible in evidence." In its ruling, the court declared: "Numerous cases are cited in support of this rule. Just when a scientific

courtroom testimony in a *Frye* state trial, the presiding judge has to determine from expert testimony that the science has such general acceptance.

Again, in a 1999 case (*Proctor v. United States*, 728 A.2d 1246, 1249), the D.C. Court acknowledged the dangers of presenting a layperson jury with expert testimony based on principles that have not been generally accepted, given the deference jurors naturally apply to experts: "Because of the authoritative quality which surrounds expert testimony, courts must reject testimony which might be given undue deference by jurors and which could thereby usurp the truthseeking function of the jury."

In another case, *Porter v. United States* (618 A.2d 619, D.C. 1992),[6] referring to statistical evidence, the Court opined: "It is almost certain that jurors would simply 'jump' to the bottom line numbers without giving any meaningful consideration to any dispute over the principles which underlie the methodology used to generate those numbers. To permit the fancy of jurors to operate in this manner is the antithesis of 'due process'.''

NRC I and NRC II

In 1989, eager to make use of the newly emerging technology of DNA profiling for the identification of suspects in a criminal case, including cold hit identifications, the FBI urged the National Research Council to carry out a study of the issue. The NRC formed the Committee on DNA Technology in Forensic Science, which issued its report in 1992. Titled *DNA Technology in Forensic Science*, and published by the National Academy Press, the report is often referred to as "NRC I". The committee's main recommendation regarding the cold hit process is given on page 124:

principle or discovery crosses the line between the experimental and demonstrable stages is difficult to define. Somewhere in this twilight zone the evidential force of the principle must be recognized, and while courts will go a long way in admitting expert testimony deduced from a well-recognized scientific courts will go a long way in admitting expert testimony deduced from a well-recognized scientific principle or discovery, the thing from which the deduction is made must be sufficiently established to have gained general acceptance in the particular field in which it belongs. We think the systolic blood pressure deception test has not yet gained such standing and scientific recognition among physiological and psychological authorities as would justify the courts in admitting expert testimony deduced from the discovery, development, and experiments thus far made." The court denied admission of the evidence. *Frye v. United States*, 293 F. 1013 (D.C. Cir. 1923)

[6] The *Porter* case was the first time the D.C. Court was faced with adjudicating the admissibility as evidence of a DNA profile match. (It was not a Cold Hit case.) The Court noted that, because a person's DNA profile is made up only of certain genetic markers and not the individual's entire unique DNA strand, it is possible for two people to coincidentally have the same profile. The Court looked to the scientific community to determine how to express the significance of a match through statistics. "Scientists calculate the possibility that the match is merely a coincidence and that the two samples did not actually come from the same person," the Court observed, and concluded, "The probability of a coincidental match is an essential part of the DNA evidence."

“The distinction between finding a match between an evidence sample and a suspect sample and finding a match between an evidence sample and one of many entries in a DNA profile databank is important. The chance of finding a match in the second case is considerably higher. … The initial match should be used as probable cause to obtain a blood sample from the suspect, but only the statistical frequency associated with the additional loci should be presented at trial (to prevent the selection bias that is inherent in searching a databank).”

For example, if the NRC I procedure were to be followed in the Jenkins case, since Jenkins was identified by a cold hit search on 8 loci, and subsequently found to have a match on all 13 CODIS loci, the prosecution could cite in court only the RMP calculated on the remaining 5 loci, namely one in one-hundred-thousand. The prosecution has repeatedly rejected this option.

In part because of the controversy the NRC I report generated among scientists regarding the methodology proposed, and in part because courts were observed to misinterpret or misapply some of the statements in the report, in 1993, Judge William Sessions, then the Director of the FBI, asked the NRC to carry out a follow-up study. A second committee was assembled, and it issued its report in 1996. Often referred to as “NRC II”, the second report, *The Evaluation of Forensic DNA Evidence*, was published by National Academy Press in 1996.

The NRC II committee’s main recommendation regarding cold hit probabilities is:

“Recommendation 5.1. When the suspect is found by a search of DNA databases, the random-match probability should be multiplied by N, the number of persons in the database.”

The statistic NRC II recommends using is generally referred to as the “database match probability”, DMP. This is an unfortunate choice of name, since the DMP is *not* a probability — although in all actual instances it is a number between 0 and 1, and it does (in my view as well as that of the NRC II committee) provide a good indication of the likelihood of getting an accidental match when a cold hit search is carried out. (The intuition is fairly clear. In a search for a match in a database of N entries, there are N chances of finding such a match.) For a true probability measure, if an event has probability 1, then it is certain to happen. However, consider a hypothetical case where a DNA database of 1,000,000 entries is searched for a profile having a RMP of 1/1,000,000. In that case, the DMP is

$$1,000,000 \ \text{x} \ 1/1,000,000 = 1$$

However, in this case the probability that the search will result in a match is not 1 but approximately 0.6312.

The committee's explanation for recommending the use of the DMP to provide a scientific measure of the accuracy of a cold hit match reads as follows:

"A special circumstance arises when the suspect is identified not by an eyewitness or by circumstantial evidence but rather by a search through a large DNA database. If the only reason that the person becomes a suspect is that his DNA profile turned up in a database, the calculations must be modified. There are several approaches, of which we discuss two. The first, advocated by the 1992 NRC report, is to base probability calculations solely on loci not used in the search. That is a sound procedure, but it wastes information, and if too many loci are used for identification of the suspect, not enough might be left for an adequate subsequent analysis. … A second procedure is to apply a simple correction: Multiply the match probability by the size of the database searched. This is the procedure we recommend." p.32.

This is essentially the same logic as I presented for my analogy with the state lottery. In the Jenkins case, the DMP associated with the original cold hit search of the (8 loci) Virginian database (containing 101,905 profiles) would be (approximately)

$$1/100{,}000{,}000 \times 100{,}000 = 1/1{,}000$$

With such a figure, the likelihood of an accidental match in a cold hit search is quite high (*cf.* the state lottery analogy). This is borne out in dramatic fashion by the Arizona study mentioned earlier, where, in a database of just 65,000 entries, 144 individuals were found with DNA profiles matching at 9 of the 13 CODIS loci.[7]

Since two reports by committees of acknowledged experts in DNA profiling technology and statistical analysis, with each report commissioned by the FBI, came out strongly against the admissibility of the RMP, one might have imagined that would be the end of the matter, and that judges in a cold hit trial would rule in favor of admitting either the RMP for loci not used in the initial identification (*à la* NRC I) or else (*à la* NRC II) the DMP but not the RMP calculated on the full match.

However, not all statisticians agree with the conclusions of the second NRC committee. Most notably, Dr. Peter Donnelly, Professor of Statistical Science at the University of Oxford, takes a view diametrically opposed to that of NRC II. In

[7] As I mentioned in an earlier footnote (page 9), the surprise in the Arizona study is not because it contradicts the mathematics — when done correctly — rather that it runs counter to the way people commonly interpret the math. I take up this issue later.

an affidavit to the Court of the District of Columbia, in connection with the Jenkins case, titled “DNA Evidence after a database hit” and dated October 3, 2004, Donnelly observes that during the preparation of the NRC II report, he had substantive discussions about the issues with four members of the committee whom he knew professionally, and goes on to say:

“I had argued, and have subsequently argued, that after a database search, the DNA evidence … is somewhat stronger than in the setting in which the suspect is identified by non-DNA evidence and subsequently found to match the profile of the crime sample. … I disagree fundamentally with the position of NRC II. Where they argue that the DNA evidence becomes less incriminating as the size of the database increases, I (and others) have argued that in fact the DNA evidence becomes stronger. … The effect of the DNA evidence after a database search is two-fold: (i) the individual on trial has a profile which matches that of the crime sample, and (ii) every other person in the database has been eliminated as a possible perpetrator because their DNA profile differs from that of the crime sample. It is the second effect, of ruling out others, which makes the DNA evidence stronger after a database search…”

Donnelly advocates using a Bayesian analysis to determine the probability of a random match, which method he outlined in a paper co-written with David Balding in 1996, titled “Evaluating DNA Profile Evidence When the Suspect is Identified Through a Database Search” (*J. Forensic Science* 603) and again in a subsequent article co-written with Richard Friedman: “DNA Database Searches And The Legal Consumption Of Scientific Evidence”, *Michigan Law Review*, 00262234, Feb99, Vol. 97, Issue 4.

The statistic generated by the Donnelly/Balding method is generally close to the RMP, although it results from a very different calculation.

The Donnelly/Balding method was considered by NRC II and expressly rejected.

We thus have a fascinating situation: two groups of highly qualified experts in statistical reasoning, each proposing a different way to calculate the likelihood that a cold hit search will identify an innocent person, and each claiming that its method is correct and the other is dead wrong.

Although Donnelly and Balding’s Bayesian approach has been accepted in courts in the UK and elsewhere in Europe, US courts have consistently taken the view (wisely, I think) that Bayesian techniques are too subtle to be understood by non-statisticians, and accordingly have never been allowed as evidence in a US court.

Adding to that concern, the FBI’s DNA Advisory Board, in its report “Statistical and Population Genetics Issues Affecting the Evaluation of the Frequency of

Occurrence of DNA Profiles Calculated From Pertinent Population Database(s)" [*Forensic Science Communications*, July 2000, Volume 2, Number 3, U.S. Department of Justice, FBI][8], wrote:

> … without the Bayesian framework, the Balding and Donnelly (1996) formulation is easily misinterpreted in a fashion unfavorable to the suspect. … [W]e continue to endorse the recommendation of the NRC II Report for the evaluation of DNA evidence from a database search.

In fact, American courts have been reluctant to rely solely on statistical evidence of any kind in determining guilt, and such use in DNA cases would be somewhat of an exception. For example, in *Brim v. Florida* (799 So. 2d 427, Fla. Dist. Ct. App. 2000), the court declared (779 So. 2d at 445 n.47):

> It should not be overlooked that courts have traditionally prohibited the use of statistical evidence to prove guilt in criminal trials. *See People v. Collins*, 68 Cal.2d 319, 66 Cal. Rptr. 497, 438 P.2d 33 (1968) (noting, "[m]athematics, a veritable sorcerer in our computerized society, while assisting the trier of fact in the search for truth, must not cast a spell over him"); *see also* Laurence H. Tribe, *Trial by Mathematics: Precision and Ritual in the Legal Process*, 84 Harv. L. Rev. 1329, 1377 (Apr. 1971) (concluding that utility of mathematical methods is greatly exaggerated, that the methods inherently conflict with other important values, and thus "the costs of attempting to integrate mathematics into the fact-finding process of a legal trial outweigh the benefits"). Thus, the admissibility of DNA statistical evidence can be viewed as a departure from the general rule.

Personally, I (together with the collective opinion of the NRC II committee) find it hard to accept Donnelly's argument, but his view does seem to establish quite clearly that the relevant scientific community (in this case statisticians) have not yet reached consensus on how best to compute the reliability metric for a cold hit, thereby ensuring that *Frye* may continue for a while longer to render inadmissible as evidence the presentation of DNA match statistics in the case of a cold hit. (I will explain why I think that two highly qualified groups of statisticians reach diametrically opposite conclusions in due course.)

In the meantime, let me bring you up to date with progress so far in the Jenkins case.

The Jenkins case

In April 2000, Raymond Jenkins was formally charged with second-degree murder while armed and in possession of a prohibited weapon, a charge that was superceded in October of the same year by one of two counts of felony murder and one count each of first-degree premeditated murder, first-degree burglary

[8] http://www.fbi.gov/hq/lab/fsc/backissu/july2000/dnastat.htm

while armed, attempted robbery while armed, and the possession of a prohibited weapon.

In March 2001, lawyers from the D.C. Public Defenders Office assigned to Jenkins' defense filed a motion to exclude the government's DNA typing results at trial, arguing that the FBI's typing methodologies were inadmissible under the standard for admission of novel scientific evidence set forth in the 1923 *Frye* ruling.

In May of that year, the government filed a response, arguing the contrary.

Three years of motions and countermotions later — during which time Jenkins' original defense counsel left the Public Defender Service and new counsel was appointed — in late March and early April 2005 the matter finally came for adjudication, when the honorably Rhonda Reid Winston granted the government's request for an evidentiary hearing as to whether it could present the RMP (more specifically, its own calculated RMP of 1 in 26 quintillion) as a generally accepted, accurate expression of the statistical significance of a cold hit match resulting from a database search. (The government added that it would not object if the defense chose to introduce the database match probability, as recommended by NRC II, but only if that figure were presented in addition to its RMP.)

The crux of the government's case was that the RMP describes a generally accepted, objective fact about DNA profiling, and that the method whereby the match was obtained is a question not of science but of procedure, and thus is not subject to exclusion under *Frye*. Citing *Porter* (another DNA profiling case, remember), lawyers representing Jenkins argued against the government's proposition, pointing out that the relevant experts disagree as to whether a database search procedure *increases*, *decreases*, or *does not effect* the strength of the match evidence, and hence disagree as to which statistical methodology is valid in a cold hit case. Both sides brought expert witnesses into court to support their respective cases, in addition to soliciting affidavits.[9]

The court heard argument from both sides on April 4, 2005, and issued a ruling the following day. Broadly speaking, the court ruled in favor of the position advocated by the defense. At the time of writing this article (November 2005), the government has filed an appeal against Judge Winston's ruling.

The statistical options

The *Jenkins* case has so far given rise to five different statistical methods for calculating the significance of a cold hit match.

[9] Including an affidavit from me.

1. Report the RMP (alone). This is the approach advocated by the government. The government cited two experts, Dr. Fred Bieber (a pathologist) and Dr. James Crow (a geneticist, and chairman of the NRC II committee), who claim that the search process has no effect on the statistical significance of a match, and hence that the RMP is a reliable indicator. While some statisticians have argued in favor of this approach, many have argued strongly against it (a view I share). The NRC II report came down firmly against any mention of the RMP in court.

2. Report the DMP (alone). This is the approach advocated by NRC II. In that report, the committee use an analogy with tossing a coin to illustrate the inapplicability of the RMP:

"[I]f we toss 20 reputedly unbiased coins once each, there is roughly one chance in a million that all 20 will show heads. According to standard statistical logic, the occurrence of this highly unlikely event would be regarded as evidence discrediting the hypothesis that the coins are unbiased. But if we repeat this experiment of 20 tosses a large enough number of times, there will be a high probability that all 20 coins will show heads in at least one experiment. In that case, an event of 20 heads would not be unusual and would not in itself be judged as evidence that the coins are unbiased. The initial identification of a suspect through a search of a DNA database is analogous…: A match by chance alone is more likely the larger the number of profiles examined."

During the Jenkins proceedings, different groups of scientists suggested that the DMP is not an accurate measure of significance for a cold hit. The government's experts argued that the DMP is a relevant statistic but interpreted NRC II (incorrectly by my reading of the NRC II report) as allowing both the RMP and the DMP to be presented in court. The Balding/Donnelly school said they believed that the DMP *under*estimates the significance of a match. A third group believed that the DMP *over*estimates the significance of a match and suggested that the NRC I method of using only confirmatory loci, not used in the initial database search, be used to calculate the RMP, which figure should then be presented to the jury as an indication of the likely accuracy of the identification.

3. Report both the RMP and the DMP. This approach is advocated by the FBI's DNA Advisory Board, which argues that both figures are "of particular interest" to the jury in a cold hit case, although it's not clear how laypersons could weigh the relative significance of the two figures, nor indeed is it at all clear that it would be right to ask them to so do, when some of the world's best statisticians are not agreed on the matter. In the original *Porter* judgment, Judge Kennedy declared that a jury should not be asked to decide scientific issues on its own.

4. Report the results of the Balding/Donnelly Bayesian approach. Roughly speaking, Balding and Donnelly believe that a cold hit carries more weight than a

purely random match and that the significance of a cold hit match *increases* as the database size goes up, arguing that an important feature of a search through a database of, say, 100,000 entries that results in a hit is that it eliminates 99,999 people from consideration.

Donnelly points out that if the search was through a database of the entire population of the world, a unique match would, barring laboratory errors, indicate certainty of guilt, and infers from this that the closer we get to such a (hypothetical?) global database, the greater becomes the accuracy of the cold hit method. Of course, if there were a global database of all living persons, there would be no need for statistics at all; a statistical analysis is required when decisions have to be made based on evidence from a small sample. Nevertheless, from a logical standpoint, Donnelly's observation does seem reasonable (at least at first blush), and hence presents a challenge to all who support the NRC II's claim that as the database size increases, so too does the likelihood of an innocent person being identified by the database search.

As mentioned earlier, Balding and Donnelly approach statistical reasoning from a strict Bayesian perspective, and I'll outline presently how such reasoning proceeds. For the moment, I note that, using Bayesian analysis to compute their own reliability statistic for a cold hit match, they arrive at a figure just slightly smaller than the RMP.

5. <u>Report the RMP calculated on confirmatory loci not considered in the initial search.</u> This is the approach advocated by NRC I, and is the only one that the entire scientific community seems to agree is reliable (though not all actively argue for its use). Several scientists have argued (in the Jenkins case and elsewhere) that this is in fact the *only* reliable procedure, and that even the NRC II method should not be used. In the Jenkins case, the government has consistently refused to follow the NRC I procedure.

Given that, at least under *Frye*, the lack of scientific consensus (which the above five positions make abundantly clear) leaves the courts unable to admit the DNA profile evidence the FBI consistently favors — namely one where a the RMP on 13 loci, or something very close to it, is presented to the jury — the only way forward might be to follow the NRC I approach, with DNA profiles based on more loci than the current CODIS system. If profiles were based on, say, 20 loci, then 10 could be used to carry out a cold hit search, leaving 10 others for confirmatory identification that could be presented in Court, using the RMP computed on 10 loci. This would give the government an admissible likelihood statistic of 1 in 10 billion, which is surely enough to convince any jury. (Of course, adoption of such a process would be expected to result in more suspects being eliminated, when the confirmatory profiling fails to produce a match. That would undoubtedly lower the FBI's conviction rate, but would clearly better serve justice.) A few years ago, this way out of the dilemma was not available, due to the limitations of DNA profiling technology, but today commercially available profiling systems exist that could allow the use of profiles on at least 21 loci. As several parties have observed (though some would dispute this), this procedure might “waste evidence”. But, by giving all parties essentially what they keep asking for, it would at least allow the process to move forward.

NRC v Balding–Donnelly

The position put forward by Prof. Peter Donnelly in the Jenkins case (and elsewhere) causes considerable conflict in many informed onlookers who are convinced by the reasoning of NRC II. (I count myself as one such.) Not just because the impressive statistical credentials that Donnelly and his like-minded colleagues bring to the debate force us to take what they say very seriously, but because, if you follow their argument with care, they seem to make a good case. And yet their conclusion appears to fly in the face of NRC II.

For example, in the Donnelly–Friedman paper “DNA Database Searches And The Legal Consumption Of Scientific Evidence” (*Michigan Law Review*, 00262234, Feb99, Vol. 97, Issue 4), the authors write:

> Though the NRC reports differ in their ultimate recommendations, their analyses of the ***database*** search problem are very similar. We believe that these analyses, and those of scholars who have supported the NRC approach, are clearly wrong. They ask the wrong question, and they fail to recognize the full import of ***evidence*** of identification based on a ***database*** search.
>
> The proper view of the situation, which we will present here, reflects a rather simple intuition. The value of a ***DNA*** match is attributable to the rarity of the profile. If the ***DNA*** of a particular person matches the crime sample, that ***evidence*** strongly supports the proposition that that person was the source of the crime sample; that is, the ***evidence*** makes that proposition appear far more probable than it did before the match was known. That other samples have been tested and found not to match does not weaken the probative value of the match, with respect to this particular proposition, which is the one of interest at the time of trial. On the contrary, this result somewhat strengthens the probative value of the match, because it eliminates some other persons as potential sources. How probable it appears that the particular person is the source depends not only on the ***DNA evidence*** but also on the other ***evidence*** in the case. If there is no other ***evidence*** pointing to him, then the proposition will not appear as likely as if there were such ***evidence*** — not because the ***DNA evidence*** is any less valuable, but because the prior probability of the proposition is so low. And ***evidence*** found after the ***DNA*** match is determined might be subject to a ground of skepticism — the possibility of suggestiveness created by the match itself — not applicable to ***evidence*** found beforehand. Thus, the probability that the defendant is the source of the crime sample may well appear less in the trawl case than in the confirmation case, but this is not because the ***DNA evidence*** itself is any weaker in the trawl case.
>
> We will now explore the reasoning that leads to these conclusions.
>
> Both NRC I and NRC II emphasized that, as the number of profiles tested

> increases, so too does the probability of finding a match with the crime sample. That is indisputably true. One can even say that the larger a ***database*** is the more likely it is that the ***database*** will yield at least one false positive result — a profile that matches the crime scene sample but that does not come from the source of that sample.(n51) But the conclusion that the NRC reports draw is that the larger a ***database*** is (up to a point) the less valuable is ***evidence*** that a ***database*** trawl yielded a single match. Here the NRC and its supporters go wrong.
>
> The proposition that the ***DNA evidence*** is offered to prove is not the broad one that the source of the crime sample is a person represented in the ***database***. Rather, it is that one particular person — the defendant in the case at hand — is the source of that sample. And the ***evidence*** bearing on this proposition is not simply that there was one match within the ***database***. Rather, it is that the ***DNA*** of that particular person — alone of all those tested — matches the crime sample.

[The emphasis in the above passage is in the original. In a moment I will argue that we should take very serious note of the words they chose to highlight. I believe it points to the fundamental issue in the matter.[10]]

Various experts in the arena of DNA profiling have made the claim that *both* the NRC committee and Balding–Donnelly are right, but that each tries to answer a different question.

For instance, in its report "Statistical and Population Genetics Issues Affecting the Evaluation of the Frequency of Occurrence of DNA Profiles Calculated From Pertinent Population Database(s)" [*Forensic Science Communications*, July 2000, Volume 2, Number 3, U.S. Department of Justice, FBI][11], the FBI's DNA Advisory Board wrote:[12]

> If we follow Balding and Donnelly (1996), the message for the investigators is that the evidence is 100,000 times more likely if the suspect is the source than if he is not. Alternatively, by the NRC II Report (1996) recommendations, the evidence is not compelling because the likelihood the profile, a priori, is/is not in the database is the same. In probabilistic terms, it is not surprising to find a matching profile in the database of size 100,000 when the profile probability is 1/100,000. Curiously, the mathematics underlying both approaches are correct, despite the apparently divergent answers. It is the foundations of the formulations that differ, and they differ substantially.

With regard to whether the two groups ask different questions, Donnelly himself writes, elsewhere in the Donnelly–Friedman article quoted from earlier:

[10] Not that I have specific information as to why the authors themselves chose to highlight certain keywords. Typographic clues may sometimes lead accidentally to fortuitous conclusions just as Cold Hit searches may sometimes lead to false identifications.

[11] http://www.fbi.gov/hq/lab/fsc/backissu/july2000/dnastat.htm

[12] The figures in the passage refer to a hypothetical example where a database of 100,000 entries is trawled for a match on a profile with an RMP of 1/100,000 and produces a single match.

> … because they fail to ask the right question, NRC II and its supporters fail to recognize the full force of proof that a single profile in a DNA database matches the crime sample.

While it is undoubtedly true that the two groups work in different frameworks of reasoning, I don't think that it's correct to say that they ask different questions. The question each tries to answer is essentially the same, namely the very one that the jury needs to know the answer to: Given an individual X who has been identified by the Cold Hit database search, what is the probability that person X is not the actual source of the DNA found at the crime scene? (Put another way, what is the probability that the DNA at the crime scene, although having the same profile as that of person X, actually comes from someone else?)

What makes the apparent conflict between NRC II and Balding–Donnelly so puzzling is that this question appears to be one that must have a single, correct answer. Surely there must be a single correct number, even if different people arrive at it using different methods, right? And if there is a correct answer, you should reach it whatever method you use. Yet NRC II and Balding–Donnelly not only have different methods of computing the answer, the numbers they arrive at are in general quite different, and moreover vary in dramatically different ways as the size of the database increases.

Given that both parties have high levels of mathematical skill, it is reasonable to assume that each is correct relative to the terms within which they are operating. If so, then it *must* be the case that they trying to do different things. In what follows, I will argue that this is indeed the case. It is not that they set out to answer different questions, but that they *interpret* the common question in two different logical frameworks, and that is why they arrive at very different answers.

Such a situation is hardly uncommon in human affairs. For instance, different frameworks of beliefs lead citizens to arrive at very different answers to the question "Who is most suitable to be the next President of the United States?" Moreover, the answer each person obtains may be entirely logical and capable of rational justification within their own belief system.

An analogous situation arises in mathematics with the classic question pertaining to Euclid's Fifth Postulate in geometry: "Given a straight line and a point not on the line, can you draw exactly one line through the point and parallel to the line?" Euclid believed the answer was yes, but in the 18th century mathematicians discovered that it all depends on what kind of geometry you are working in, an observation that became more significant in general terms when physicists realized in the 20th century that the geometry of the universe we live in is in fact one of the *non*-Euclidean varieties. Euclidian geometry is "correct" for some real world applications; for other applications, you need to work in a different geometric framework.

None of the different kinds of geometries is intrinsically “right” or “wrong” from a mathematical perspective. The issue is which geometry best meets the needs of the application of concern. In the case of applying geometry to the physical universe, that issue is decided by the physics itself. If my analogy between the situation in geometry and the current DNA profile cold hit debate is sufficiently valid,[13] then in the cold hit case, presumably it will (in the final analysis) be up to the courts to make that determination — although one would hope that they do so at a federal level and only with considerable input from experts.

A significant, and perhaps surprising feature of the different frameworks adopted by the two main parties in the Cold Hit debate (NRC II and Balding/Donnelly), is that they mean different things by the word *probability*!

In that regard they are not alone.

What exactly does a numerical probability tell us?

Many people have considerable difficulty reasoning about probabilities. In my experience, the vast majority of times when mathematically able people have problems reasoning about probabilities it is because they unconsciously confuse two very different notions, both of which are called “probability”. Indeed, many people are not even aware that the word has two quite different (though consistent) meanings.

For the kinds of example that teachers and professors typically use to introduce students to probability theory, the answer to the question asked by the section heading seems clear-cut. If you toss a fair coin, they say, the probability that it will come down heads is 0.5 (or 50%).[14] What this means, they go on to say, is that, if you tossed the coin, say, 100 times, then roughly 50 times it would come down heads; if you tossed it 1,000 times, it would come down heads roughly 500 times; 10,000 tosses and heads would result roughly 5,000 times; and so forth. The actual numbers may vary each time you repeat the entire process, but in the long run you will find that roughly half the time the coin will land on heads. This can be expressed by saying that the probability of getting heads is 1/2, or 0.5.

Similarly, if you roll a fair die repeatedly, you will discover that it lands on 3 roughly 1/6 of the time, so the probability of rolling a 3 is 1/6.

[13] And there is a vagueness here since we are talking about analogies.

[14] Actually, as the mathematician Persi Diaconis demonstrated not long ago, it’s not exactly 0.5; the physical constraints of tossing an actual coin result in roughly a 0.51 probability that it will land the same way up as it starts. But I’ll ignore that wrinkle for the purposes of this explanation.

In general, if an action A is performed repeatedly, the probability of getting the outcome E is calculated by taking the number of different way E can arise and dividing by the total number of different outcomes that can arise from A. Thus, the probability that rolling a fair die will result in getting an even number is given by calculating the number of ways you can get an even number (namely 3, since each of 2, 4, and 6 is a possible even number outcome) and dividing by the total number of possible outcomes (namely 6, since each of 1, 2, 3, 4, 5, 6 is a possible outcome). The answer, then, is 3 divided by 6, or 0.5.

Most instructors actually get the students to carry out coin tossing and dice rolling experiments for themselves to help them develop a good sense of how numerical probabilities arise.

Notice that the probability — the number — is assigned to a single event, not the repetition of the action. In the case of rolling a die, the probability of 0.5 that the outcome will be even is a feature of the single action of rolling the die (once). It tells you something about how that single action is likely to turn out. Nevertheless, it derives from the behavior that will arise over many repetitions, and it is only by repeating the action many times that you are likely to observe the pattern of outcomes that the probability figure captures. The probability of a particular outcome of an action is a feature of that single outcome that manifests itself only when the action is performed repeatedly.

Probability is, then, an empirical notion. You can test it by experiment. At least, the kind of probability you get by looking at coin tossing, dice rolling, and similar activities is an empirical notion. What causes confusion for many people is that mathematicians were not content to leave the matter of trying to quantify outcomes in the realm of games of chance — the purpose for which (mathematical) probability theory was first developed in the seventeenth century. They took it out into the everyday world — a decidedly less precise and self-contained environment. And when they did so, it took on a quite different meaning.

By way of an illustration of how probability theory can be applied in the everyday world, consider the following scenario. Suppose you come to know that I have a daughter who works at Google; perhaps you meet her. I then tell you that I have two children. This is all you know about my family. What do you judge to be the likelihood (dare I say, the probability?) that I have two daughters? (For the purposes of this example, we'll assume that boys and girls are born with exactly 50% likelihood.)

If you are like many people, you will argue as follows. "I know Devlin has one daughter. His other child is as likely to be a boy as a girl. Therefore the probability that he has two daughters is 1/2 (i.e., 0.5, or 50%)."

That reasoning is fallacious. If you reason correctly, the probability to assign to my having two daughters is 1/3. Here is the valid reasoning. In order of birth, the gender of my children could be B-B, B-G, G-B, G-G. Since you know that one of my children is a girl, you know that the first possibility listed here does not arise. That is, you know that the gender of my children in order of birth is one of B-G, G-B, G-G. Of these three possibilities, in two of them I have one child of each gender, and in only one do I have two daughters. So your assessment of the likelihood of my having two daughters is 1 out of 3, i.e., probability 1/3.

But even if you figure it out correctly, what exactly is the significance of that 1/3 figure? As a matter of fact, I *do* have two children and one of my children *is* a daughter who works at Google. Does anyone believe that I live in some strange quantum-indeterminate world in which my other child is 1/3 daughter and 2/3 son? Surely not. Rather, that 1/3 probability is a measure not of the way the world is, but of *your knowledge* of the world (to be specific, your knowledge about my family).

As it happens, I do have two daughters, and not surprisingly, I am aware of the gender of my children. So if you asked *me* what probability *I* myself would assign to my having two daughters, I would say probability 1. This is a different answer from yours, reflecting the fact that we have different knowledge about the world.

The probability of 1/3 you (should) put on your knowledge about my two children being girls is a feature of your knowledge. The probability of 1 that I put on my knowledge about my two children being girls is a feature of my knowledge. Neither of these probabilities is an objective feature of the world (though my knowledge happens to agree with the world in this case). There is an objective probability that is associated with the event that both my daughters are girls, and that probability is 1. Obviously, when the event has already taken place, the objective probability of that event can only be 0 or 1. Probabilities attached to the world — not to someone's knowledge of the world — can only be other than 0 or 1 when the event has not yet taken place.

The concept of probability you get from looking at coin tossing, dice rolling, and so forth is generally referred to as "objective probability" or "frequentist probability". It applies when there is an action, having a fixed number of possible outcomes, that can be repeated indefinitely. It is an empirical notion, that you can check by carrying out experiments.

The numerical measure people assign *to their knowledge* of some event is often referred to as "epistemic probability" or "subjective probability".[15] It quantifies an individual's *knowledge of the event*, not the event itself. Different people can

[15] As used historically, the terms "epistemic probability" and "subjective probability" are not completely synonymous, but the distinction is outside the scope of this article.

assign different probabilities to their individual knowledge of the same event. The probability you assign to an event depends on your prior knowledge of the event, and can change when you acquire new information about it.

An objective (or frequentist) probability *can* be viewed as a subjective probability. (A Bayesian would say it *has* to be so viewed.) For instance, the probability of 1/2 that I assign to the possibility of getting a head when I toss a fair coin ten minutes from now is, when thought of as a measure of my current knowledge about a future event, a subjective probability according to the definition just given. (Clearly, when we quantify our information about a future occurrence of a repeatable action, where the frequentist notion of probability applies, we should assign the frequentist value.)

To drive home the crucial fact that there is a distinction here, imagine that the above coin toss has already taken place. That is, I have just tossed a coin, and seen that it came up tails, but you cannot see it and do not yet know the outcome. What would *you* say is the probability that the coin came up heads? The only rational answer you could give is 0.5. And yet the toss has already occurred. Either it *is* heads or it *is* tails. *In the world*, the probability that the coin is heads up is 0 and the probability that it is tails up is 1. Those are also the probabilities I would assign to *my knowledge* of the coin toss. How come that you calculate a different probability? Because what you are quantifying with your calculation is *your knowledge* of the world.

I am sure that a confusion between the objective/frequentist and subjective/ epistemic notions of probability is what lies behind the problem many people have in understanding the reasoning of the notorious Monty Hall problem. That problem is posed to *appear* to be about a physical situation (where a prize is hidden) but in fact it is not; it's about your individual knowledge of that situation, and how that knowledge changes as you receive additional information.

It is well worth going through this problem in some detail, as I believe the (fundamental) confusions it highlights are precisely what is causing much of the debate about Cold Hit probabilities.

The Monty Hall problem

In the 1960s, there was a popular weekly US television quiz show called *Let's Make a Deal.* Each week, at a certain point in the program, the host, Monty Hall, would present the contestant with three doors. Behind one door was a substantial prize; behind the others there was nothing. Monty asked the contestant to pick a door. Clearly, the probability of the contestant choosing the door with the prize was 1 in 3 (i.e., 1/3). So far so good.

Now comes the twist. Instead of simply opening the chosen door to reveal what lay behind, Monty would open one of the two doors the contestant had not chosen, revealing that it did not hide the prize. (Since Monty knew where the prize was, he could always do this.) He then offered the contestant the opportunity of either sticking with their original choice of door, or else switching it for the other unopened door.[16]

The question now is, does it make any difference to the contestant's chances of winning to switch, or might they just as well stick with the door they have already chosen?

When they first meet this problem, many people think that it makes no difference if they switch. They reason like this: "There are two unopened doors. The prize is behind one of them. The probability that it is behind the one I picked is 1/3, the probability that it is behind the one I didn't is the same, that is, it is also 1/3, so it makes no difference if I switch."

A common variant is for people to think that the two probabilities are not 1/3 and 1/3, but 1/2 and 1/2. Again, the intuition is that they are faced with two equally likely outcomes, but instead of regarding them as two equal choices that remain from an initial range of three options, they view the choice facing them as a completely new situation. This accords with the experience we all have in games like coin tossing and dice rolling, where each new toss or roll is a brand new event, totally uninfluenced by any previous tosses or rolls.

Surprising though it may seem at first, however, either variant of this reasoning is wrong. Switching actually *doubles* the contestant's chance of winning. The odds go up from the original 1/3 for the chosen door, to 2/3 that the *other* unopened door hides the prize.

Yet again, a grounding experience in probabilities in gambling games leads people astray when it comes to reasoning about knowledge.

There are several ways to explain what is going on here. Here is what I think is the simplest account.

Imagine you are the contestant. Suppose the doors are labeled A, B, and C. Let's assume you (the contestant) initially pick door A. The probability that the prize is behind door A is 1/3. That means that the probability it is behind one of the other two doors (B or C) is 2/3. Monty now opens one of the doors B and C to reveal

[16] As the game was actually played, some weeks Monty would simply let the contestant open their chosen door. The hypothetical version of the game described here, where Monty always opens the door and makes the "switch or stick" offer, is the one typically analyzed in statistics classes.

that there is no prize there. Let’s suppose he opens door C. (Notice that he can always do this because he knows where the prize is located.) You (the contestant) now have two relevant pieces of information:

1. The probability that the prize is behind door B or C (i.e., not behind door A) is 2/3.

2. The prize is not behind door C.

Combining these two pieces of information, you conclude that the probability that the prize is behind door B is 2/3.

Hence you would be wise to switch from the original choice of door A (probability of winning 1/3) to door B (probability 2/3).

Now, experience tells me that if you haven’t come across this problem before, there is a good chance that the above explanation fails to convince you.

The instinct that compels people to reject the above explanation is, I think, a deep-rooted sense that probabilities are fixed. That is true in the case of frequentist probabilities — probabilities about the way the world is. But that is not what the Monty Hall problem is about. The game has already been set up. The prize already is behind exactly one of the doors. There is no uncertainty. Moreover, Monty knows where the prize is. He too has no uncertainty. But you, the game show contestant, do not have that certain knowledge. You are having to reason about *your* knowledge. And that can change as you acquire more information. It is because the acquisition of information changes the probabilities associated with different choices that we often seek information prior to making an important decision. Acquiring more information about our options can reduce the number of possibilities and narrow the odds.

When you are the game show contestant and Monty opens his door and shows you that there is no prize behind it, he thereby injects a crucial piece of information into the situation. Information that you can take advantage of to improve your odds of winning the grand prize. By opening his door, Monty is in effect saying to you:

“There are two doors you did not choose, and the probability that the prize is behind one of them is 2/3. I’ll help you by using my knowledge of where the prize is to open one of those two doors to show you that it does not hide the prize. You can now take advantage of this additional information. Your choice of door A has a chance of 1 in 3 of being the winner. I have not changed that. But by eliminating door C, I have shown you that the probability that door B hides the prize is 2/3.”

Still not convinced? Some people who have trouble with the above explanation find it gets clearer when the problem is generalized to 100 doors. As the contestant, you choose one door. You will agree, I think, that you are highly likely to lose if you open that door. The chances are highly likely (in fact 99/100) that the prize is behind one of the 99 remaining doors. Monty now opens 98 of those other doors and none of them hides the prize. There are now just two remaining possibilities: either your initial choice was right or else the prize is behind the remaining door that you did not choose and Monty did not open. Now, you began by being pretty sure you had little chance of being right — just 1/100 in fact. Are you now saying that Monty's action of opening 98 doors to reveal no prize (carefully avoiding opening the door that hides the prize, if it is behind one of the 99 you did not choose) has increased to 1/2 your odds of winning with your original choice? Surely not. In which case, the odds are high — 99/100 to be exact — that the prize lies behind that one unchosen door that Monty did not open. You should *definitely* switch. You'd be crazy not to!

Let me make one last attempt at an explanation. Back to the three door version now. When Monty has opened one of the three doors and shown you there is no prize behind, and then offers you the opportunity to switch, he is in effect offering you a *two-for-one* switch. You originally picked door A. He is now saying, in effect, "Would you like to swap door A for *two* doors, B and C? Oh, and by the way, before you make this two-for-one swap I'll open one of those two doors for you, one without a prize behind it."

In effect, then, when Monty opens door C, the attractive 2/3 odds that the prize is behind door B or C are shifted to door B alone.

So much for the mathematical explanations. But at least as fascinating as the mathematics, to my mind, is the psychology that goes along with the problem. Not only do many people get the wrong answer initially (believing that switching makes no difference), but a substantial proportion of them are unable to escape from their initial confusion and grasp any of the different explanations that are available (some of which I gave above).

On those occasions when I have entered into some correspondence with laypersons, and there have been many of them over the years, I have always prefaced my explanations and comments by observing that this problem is notoriously problematic, that it has been used for years as a standard example in university probability courses to demonstrate how easily we can be misled about probabilities, and that it is important to pay attention to every aspect of the way Monty presents the challenge. Nevertheless, I regularly encounter people who are unable to break free of their initial conception of the problem, and thus unable to follow any of the explanations of the correct answer.

Indeed, some individuals I have encountered are so convinced that their (faulty) reasoning is correct, that when I try to explain where they are going wrong, they become passionate, sometimes angry, and occasionally even abusive. Abusive over a math problem? Why is it that some people feel that their ability to compute a game show probability is something so important that they become passionately attached to their reasoning, and resist all attempts by me and others to explain their error? On a human level, what exactly is going on here?

First, it has to be said that the game scenario is a very cunning one, cleverly designed to lead the unsuspecting player astray. It gives the impression that, after Monty has opened one door, the contestant is being offered a choice between two doors, each of which is equally likely to lead to the prize. That would be the case if nothing had occurred to give the contestant new information. But Monty's opening of a door does yield new information. That new information is primarily about the two doors not chosen. Hence the two unopened doors that the contestant faces at the end are not equally likely to lead to the prize. They have different histories. And those different histories lead to different probabilities.

That explains why very smart people, including many good mathematicians when they first encounter the problem, are misled. But why the passion with which many continue to hold on to their false conclusion? I have not encountered such a reaction when I have corrected students' mistakes in algebra or calculus.

I think the reason the Monty Hall problem raises people's ire is because a basic ability to estimate likelihoods of events is important in everyday life. We make (loose, and generally non-numeric) probability estimates all the time. Our ability to do this says something about our rationality — our capacity to live successful lives — and our rationality is one of the distinctive features of being human. The degree of our ability to reason can be regarded as a measure of "how good" a person we are. It can thus be a matter of pride, something to be defended. (Few people see calculus the same way, I regret to say, but then, unlike reasoning about the likelihoods of future events, there have been few moments in the long history of human evolution where skill in calculus has offered any survival value.)

Many people have trouble with the Monty Hall reasoning presented above because they cannot accept that the probability attached to a certain door changes. They argue that if the probability that the prize is behind any one door is 1/3 at the start, then, because no one has moved the prize, it must still be 1/3. Hence it makes no difference whether they stick with their original choice or switch. The probability that it is behind *either* unopened door is 1/3 — they face the same odds whether they stick or switch.

What they are doing is confusing two separate issues: how things are in the world and what they know about the world. It is true that, since no one moves the prize or otherwise alters the physical arrangement of the game, the "probabilities-in-the-world" do not change. But those probabilities are not 1/3, 1/3, 1/3: they are 0 for the two doors where there is no prize and 1 for the door where it is. (In other words, there is no uncertainty at all; the matter has been settled.) But the contestant's *knowledge* of the situation most definitely does change. Monty's opening one door provides the contestant with new information, information that on this occasion changes the contestant's-knowledge-probabilities from 1/3, 1/3, 1/3 to 1/3, 2/3, 0 (assuming the contestant originally picks door A and Monty opens door C).

The reasoning that the contestant needs to go through here is not about the way the world is — although a superficial reading might suggest that he or she is so doing — rather it is about the *information* the contestant has about the world. As such, it is an example of Bayesian reasoning. At least, it would be if I had presented the argument a bit differently. I'll come back to that issue momentarily. For now, it's more accurate to say that, by focusing on what the contestant knows and how the probabilities change when he or she acquires new information, the reasoning I gave was at least in the *philosophical spirit* of the Bayesian approach.

Bayesian reasoning

In Bayesian reasoning, probabilities are attached not to states of the world but to statements (or propositions) about the world. You begin with an initial assessment of the probability attached to a statement (in the Monty Hall example it's the statement that the prize is behind the unchosen door B, say, to which you assign an initial, or *prior* probability of 1/3). Then you modify that probability assessment based on the new information you receive (in this case, the opening of door C to reveal that there is no prize behind it) to give a revised, or *posterior* probability for that statement. (In this case the posterior probability you attach to the statement that the prize is behind door B is 2/3.)

There is a specific and precise rule that tells you how to modify the prior probability to give the posterior probability: Bayes' Theorem. (I did not use that rule in my solution to Monty Hall above — I'll come back to that point in due course.) The exact formulation of Bayes' Theorem and exactly how it is used are a bit technical, and not important to the main thrust of this article, so I'll relegate a discussion to an appendix. The crucial thing to know is the general framework under which Bayesian reasoning takes place:

1. Bayesian reasoning is not about the world *per se*, but about the knowledge the reasoner has about the world.
2. Thus, in Bayesian reasoning, probabilities are attached not to events but to statements (or propositions), S.

3. Bayesian reasoning is a formal reasoning procedure whereby an individual can reason about the probability that he or she rationally attaches to the likely validity of a certain statement S.
4. The reasoner begins with some initial probability *p* that S — the prior probability.
5. The reasoner acquires some new information E, and on the basis of E is able to modify *p* to give a new probability estimate *q* for S — the posterior probability.
6. Bayes' Theorem provides a precise formula, dependent on E, for going from *p* to *q*.
7. If further information F comes in, the reasoner may repeat this process, starting with *q* as the prior probability and applying Bayes' Theorem using the formula that comes from F.
8. And so on. When the process is completed, the reasoner has arrived at a new (and hopefully more reliable) probability that S, which takes account of all the evidence obtained.

Notice that in 2 above, the statement S will most likely be about the world. Thus, in reasoning about the likely validity of S, the reasoner is highly likely to talk about the world. But the probabilities the reasoner computes are all attached to the possible validity of S. This may seem like a trivial point, and for very simple scenarios (such as Monty Hall, where the scenario is simple, even if the reasoning can be highly problematic for many people) it may be hard to follow the distinction I am making, but this is what lies behind the debate — now over a decade long — between the Bayesians such as Balding and Donnelly and the advocates of the NRC procedure(s) regarding how to calculate the probability that a Cold Hit match is a false positive. The NRC supporters calculate probabilities attached to the world — the basic question they ask is "What might the world be like?" The Bayesians calculate probabilities attached to their knowledge of the world — to the *evidence* — and the basic question they ask is "How accurately do I know what I know?" (Now you know why I was so struck by the fact that in the article I excerpted above, Donnelly talked the whole time about *evidence* and highlighted every single occurrence of the word "evidence".)[17]

Now, both approaches are mathematically correct. That means that mathematics (and mathematicians) are not going to be able to rule between them with regards to which approach should or should not be used in court proceedings. Ultimately,

[17] Although the distinction between objective/frequentist and epistemic/subjective probabilities is often discussed in more specialized works on probability theory (particularly works focusing on the philosophical aspects of probability), I have not seen reference to the distinction in terms of "probabilities in the world" and "probabilities about knowledge in the head" that I am doing here. After many years of experience discussing confusions about probabilities with lay persons, I believe that making the distinction the way I do may help many to clarify their understanding of the issues.

that will be up to the courts themselves to decide. (I'll come back later to look at what that decision might amount to in terms of how mathematics applies to the world. Mathematics might not be able to decide the issue, but I think it can provide some useful background information relevant to making such a decision.)

What mathematics (and mathematicians) can be definite — and *emphatic* — about, however, is that the one thing that absolutely must not be allowed to happen is that a court admits evidentiary reasoning that combines the two approaches, arguing one way for one fact, the other way for the next, and so on. The two approaches are not compatible. Try to combine them and the resulting reasoning will not be logically sound. Once one approach is chosen, the other must be ruled out of court. It's like crossing the Atlantic by airplane. Before you start, you have a choice between United Airlines and Air France. Either will get you there safely, although the two carriers come from different cultures and speak different languages. But once you have taken off, if you attempt to switch planes, disaster will follow immediately.

One immediate difficulty in adopting a Bayesian approach is that, for anyone who first learned about probability calculations by considering coin tossing or dice rolling examples, or other gambling games — and that is most of us — it can be very difficult to break away from that mindset and adopt a strictly Bayesian approach.

For example, many people are never able to follow any of the arguments I gave above for resolving the Monty Hall puzzle, so they resort to performing a simulation, where they pay the game many times using the switch strategy and an equal number of times using the stick strategy, and keeping track of the win-loss record in each run. (The simulation can be played physically, with a friend putting a nickel under one of three upturned eggcups each time and acting as Monty Hall, or by writing a computer simulation, or, these days, by navigating to one of a number of Websites that have Monty Hall simulators on them.) When the skeptics see that they win roughly 2/3 of the games using the switch strategy and roughly 1/3 of the games when they follow the stick strategy, they finally acknowledge that switching does indeed double the chances of winning. Individuals who have to resort to a simulation to resolve the matter seem to have a genuine block to being able to adopt a Bayesian approach and reason about knowledge; they feel they *have* to reason directly about the world.

Monty Hall with a twist

Turning now to my point that it is important not to mix Bayesian reasoning with frequentist type arguments, consider a slightly modified version of the Monty Hall game. In this variant, after you (the contestant) have chosen your door (door A, say), Monty asks *another contestant* to open one of the other two doors. That contestant, who like you has no idea where the prize is, opens one at random, let us say, door C, and you both see that there is no prize there. As in the original game, Monty now asks you if you want to switch or stick with your original choice. What is your best strategy?

The difference this time is that the choice of the door that was opened by the other contestant was purely random, whereas in the original Monty Hall game, Monty knew where the prize was and was able to use that knowledge in order to ensure that he never opened a door to reveal the prize. This time, however, with the revised procedure, there was a chance that the other contestant might have opened the door with the prize behind it.

So, what would you do? If you adopt the reasoning I gave earlier for the original Monty game, you will arrive at the same conclusion as before, namely that you should switch from door A, and that exactly as before, if you do so you will double your likelihood of winning. Why? Well, you will reason, you modified the probability of the prize being behind door B from 1/3 to 2/3 because you acquired the new information that there was definitely no prize behind door C. It does not matter, you will say, whether door C was opened (to reveal no prize) by deliberate choice or randomly. Either way, you get the same crucial piece of information: *that the prize is not behind door C*. The argument (about what you know and what you learn) remains valid. Doesn't it?

Well, no, as a matter of fact it doesn't. The problem was, although I took a Bayesian approach to solving the original Monty puzzle, I did not use *Bayesian reasoning*, which requires that you use Bayes' theorem to revise probabilities. Instead, I used everyday, run-of-the-mill, logical reasoning. As any Bayesian will tell you, in revising probabilities as a result of the acquisition of new information, it is important to know where that information came from, or at least to know the probability that it would have arisen under the prevailing circumstances. In the original Monty problem, Monty knows from the start where the prize is, and he uses that knowledge in order to always open a door that does not hide a prize. Moreover, you, the contestant, know that Monty plays this way. (This is crucial to your reasoning, although you probably never realized that fact.) Bayes' theorem explicitly takes account of the probability that the new information would have arisen in the given circumstance. In the argument I gave, I did not do that, since Monty's strategy effectively took care of that requirement for me.

In the modified Monty game, where the door is opened randomly by another contestant, the likelihood that you obtain the evidence you do is different. Apply Bayes' theorem to this game and you reach the conclusion that the probabilities of the prize being behind door A or door B (after door C has been opened and shown not to lead to the prize) are equal.

The problem with the solution to the revised problem that I gave a moment ago is that, when I argued

> Either way, you get the same crucial piece of information: *that the prize is not behind door B*. The argument (about what you know and what you learn) remains valid.

I was not reasoning about the information but *about the world*. In the world, as a matter of empirical fact, when door C was opened it turned out that there was no prize.

As a result of confusing reasoning about the world with reasoning about my information about the world, I arrived at an incorrect answer.

This example highlights the point I made earlier that it is crucial not to combine frequentist reasoning about the world with Bayesian reasoning about your knowledge of the world.

It also illustrates just how subtle the distinction can be, and how easy it can be to slip up. To repeat my earlier point, if you adopt a Bayesian approach, you have to do so totally. For anyone who learned about probabilities by looking at gambling games, and that is very likely all of us, that can be extremely hard to do.

If you try to reason through the modified Monty game using a frequentist-style approach, focusing not on the information you have at each stage and what probabilities you can attach to what you know, but instead attaching probabilities to the different ways the game might come out, then you will reach the correct answer, namely that in the modified game it makes no difference to you whether you switch or stick. And if you simulate the modified game many times, you will indeed find that it makes no difference which strategy you adopt, always switch or always stick; you will win roughly 1/3 of the plays. (With the modified game, the other contestant will also win roughly 1/3 of the plays, by opening the door to claim the prize before you get the option to switch.)

Here, briefly, is the argument:

1. You choose one door, say, door A. The probability that the prize is there is 1/3. (i.e., you will win in roughly 1 out of 3 games)

2. The probability that the prize is behind one of door B and door C is 2/3.

3. The other contestant has a choice between door B and door C. The odds she faces are equal. Assume she picks door C. The probability that she wins is 1/2 x 2/3 = 1/3.

4. The probability that she loses is likewise 1/2 x 2/3 = 1/3. And that’s the probability that you win if you switch. Exactly the same as if you did not.

The choice facing the Jenkins court

At present, the issue that has kept the DNA cold hit identification evidence out of the court in the Jenkins case is the absence of consensus in the scientific community regarding how to calculate the statistical significance of a cold hit match. Notwithstanding the introduction into the evidentiary hearings process of no less than five different ways to proceed, there really is only one disagreement of any real *scientific* substance, and that is between, on the one hand the Balding/Donnelly Bayesian camp, and on the other hand those who are in general agreement with the positions espoused by the two NRC committees, in their reports NRC I and NRC II. Accordingly, it is on that specific two-party dispute that I shall focus my attention. Resolve that, and a "scientific consensus" acceptable to the court will surely follow quickly.

Personally, as someone trained in formal logic, having a lifelong professional interest in forms of reasoning, I find the Bayesian school of thought a particularly attractive one with considerable merit. But the method is not without its difficulties, and one of them in particular is considerable. (I'll get to that momentarily.)

In the final analysis, the disagreement between the NRC camp and the Balding/Donnelly Bayesian camp presents the *Jenkins* court (and any other court trying to decide a cold hit case) with two distinct ways to proceed with regard to the use (and hence admissibility) of statistics.

- Is it the court's job to reason (in rational fashion) *about the world(s)* of Dennis Dolinger and Raymond Jenkins, using statistical methods to try to put numerical measures on the various things that may or may not have happened *in the world*? This is the philosophy behind the NRC approach.

- Or is it the court's job to focus not on the world but *the evidence* before it, and use statistical methods to try to put numerical measures on the reliability of that *evidence*? This is the Bayesian philosophy.

Arguably the court seeks to do both, but as I illustrated with my fairly lengthy discussion of the Monty Hall problem (and variants thereof), in cases where statistical inference is involved, this is not an option: the court must choose one means and stick rigorously to it. Try to combine the two and things can go badly wrong.

Donnelly advises us that the focus should be exclusively on the *evidence* — this is precisely the Bayesians' position. In his various writings, some of which I quoted earlier, Donnelly has explained how the application of Bayes' theorem may be used to modify the probability of guilt after a cold hit match (such as in the Jenkins case) based on the evidence from the DNA profile.

Now, it is the very essence of the Bayesian approach — and crucial to its success — that the entire reasoning process focuses exclusively on the *evidence*. But suppose that, *by virtue of the way the world happened to be at the time*, Jenkins was innocent of the Dolinger murder. Suppose that, as a matter not of theory or statistics but of *plain fact in the world*, that he was like the lucky jackpot winner, and that he became a suspect *solely by virtue* of the investigating agencies having searched through a sufficient number of DNA profiles that the number of profiles searched was of the same order of magnitude as the degree of rarity of that profile in the relevant general population, thereby making it highly likely that a match would be found?

It is precisely this possibility that the two NRC committees sought to account for. It is precisely because, as indicated by the application not of sophisticated statistical techniques but of simple counting, there is a very real possibility that Jenkins is a suspect purely because the authorities looked at sufficiently many DNA profiles that they were bound, sooner or later, to find *someone* whose profile matched, that the two NRC reports stipulate that the search process itself does not amount to *evidence*, and any evidence submitted to secure a conviction must be found elsewhere, either by a DNA profile comparison carried out elsewhere on Jenkins' genome or else by alternative means altogether.

Moving on, what of Donnelly's point that there is an inconsistency between the NRC II claim that as the size of the DNA database searched increases, the evidentiary value of a match decreases, and the fact that if there were a database of every living person on earth, then a match obtained by such a search would be conclusive proof of identity.

Well, first it should be noted that a key feature of Donnelly's argument is that the database search yields a *unique* match. For instance, in Donnelly & Friedman's article "DNA Database Searches And The Legal Consumption Of Scientific Evidence" [*Michigan Law Review*, 00262234, Feb99, Vol. 97, Issue 4] we read:[18]

> Now consider in addition the fact that other samples have in fact been tested and found not to match the crime sample. With respect to the precise proposition at issue — that Matcher is the source of the crime sample — this fact can only enhance the probative value of the ***DNA evidence***. One reason for this is that the additional information that a significant number of persons have been tested and found not to match the crime sample can only make the profile of that sample appear rarer than it did absent that information. This factor will almost always be of negligible importance, and we will put it aside for further analysis. Potentially more importantly, a number of people other than the defendant who previously appeared to be possible sources of the crime sample

[18] The authors refer to the individual identified by the cold hit search as Matcher.

> have now been eliminated, thus making each of the remaining possibilities somewhat more probable. Assuming, as is usually the case, that the size of the ***database*** is very small in comparison to the suspect population, this effect too will be negligible, but as the size of the ***database*** increases in comparison to that population, the effect becomes dominant. If the ***database*** includes the entire suspect population, then the existence of only one match points the finger without doubt (assuming accurate testing) at the person so identified.

In the Jenkins case, there was a unique match, but that search was done on just 8 loci, and the recent Arizona study I mentioned earlier turned up multiple pairs of individuals whose profile matched on 8 loci among a very small database, so it seems possible that the uniqueness in the Jenkins case is just happenstance.[19] With profiles based on 13 (or more) loci, given the sizes of databases in use today, at most one match is highly likely. But with extremely large databases that approach the population of the entire world — the size of database Donnelly relies on to try to refute the NRC II position — it is likely that multiple matches will be the norm.

But let’s put that issue aside for now. The Bayesian school would have the court reason according to strict Bayesian methodology, focusing solely on the evidence. Let us for the moment grant them that opportunity. In order to even begin the Bayesian belief revision process (i.e., in order to start applying Bayes’ theorem) leading toward establishing the guilt of Raymond Jenkins, you have to assign an initial probability (the prior probability) p to the proposition that Jenkins is the person who murdered Dennis Dolinger. Having done that, you take the evidence — including any available (i.e., admissible) evidence about the DNA profile match — and use it with Bayes’ theorem in order to revise p (i.e., to obtain a posterior probability q).

Well, Jenkins came under suspicion in the first place solely as a result of a cold hit DNA profile database search based on 8 loci. Thus an eminently justifiable value to take for p is the RMP associated with 8-loci searches, namely $p = 1/100,000,000$.

But making use of the RMP for the original match in order to determine the prior probability means it cannot be used as Bayesian “evidence” during the subsequent Bayesian inference process. Hence, in the subsequent applications of Bayes’ theorem, the only DNA matching evidence that may be used is for matches on additional loci, of which there is a current 5-loci match (though perhaps additional testing could be done on additional loci, should this be thought desirable).

[19] Lest it be assumed that I am here relying on multiple instances of unfortunate instances of happenstance for Jenkins, let me point out that this is the same happenstance that put him under suspicion in the first place. Everything else follows from that one initial database match, which may or may not have been accidental.

The prosecution in the Jenkins' case, however, wishes to use the entire 13-loci match as evidence. But, Jenkins was a suspect in the first place only because of the database match, so if that match (and its associated RMP of 1 in ten trillion) is to be treated as evidence in the Bayesian inference process, then *no probability* associated with that database search can be taken for the initial prior probability *p*, nor indeed be used in any way in order to determine *p*.

So how then do you determine *p*? Since, and this is not disputed by anyone involved in the case, Jenkins was *not* a suspect prior to the database search, the only mathematically justifiable value for *p* that remains is *p* = 0.

[Note added February 11, 2007: This paper is a work in progress, and the argument presented here is not yet fully worked out. In fact, it is little more than a note to myself to come back to the issue when I have the time. However, a number of lawyers have read this draft and have queried me about the issue I raise, so I need to add a caution that this is a part of the paper that is not yet completed. I believe that the question of the priors lies at the heart of the current disagreement between the NRC II committee and Balding-Donnelly. Although *p* = 0 is, I believe the only prior justifiable on mathematical grounds alone, all that really tells us is that the "correct" prior must be determined using additional criteria. But that does not mean the mathematical considerations can simply be thrown out of the window. Bayesian inference is a powerful mathematical tool, and all aspects of its use require careful consideration from a mathematical perspective.]

Now, you can apply Bayes' theorem as often as you like, but if you start with a prior probability of 0, then you will get 0 at every stage. You will continue to get a probative value of 0 after incorporating (via Bayes' theorem) *any evidence* obtained from the DNA profile match the FBI obtained on the 5 CODIS loci not used in the original cold hit search and after incorporating *any evidence* obtained from any other sources. That would of course make life extremely easy for the Defense Counsel in the case, but it is unlikely that the prosecution would want to go down that path.

Donnelly is of course aware of this possibility, and says that at the start of the murder investigation, *everyone* is a suspect, and hence that *everyone in the world* has a *p* value that is small but not zero. He writes:[20]

> Even if at any stage of the investigation or trial there is not yet any ***evidence*** pointing to Matcher, he, like everybody else in the world who had not been eliminated as a possibility, is in fact a suspect in the limited sense that he is possibly the source of the crime sample. Thus, an investigator or factfinder asked to do so might assign a probability to the proposition that Matcher is the source.

You might ask yourself by what scientific means "an investigator or factfinder asked to do so might assign a probability to the proposition that Matcher is the source."

[20] P. Donnelly and D. Friedman, "DNA Database Searches And The Legal Consumption Of Scientific Evidence" [*Michigan Law Review*, 00262234, Feb99, Vol. 97, Issue 4]

(Incidentally, Donnelly needs to jump all the way to everyone in the world here because it is only when there is a world database that his argument against NRC II becomes effective.)

But Donnelly's position that at the start of a criminal investigation *every living person in the world* is a suspect, not only flies in the face of our widely accepted principle of innocent until proved guilty, but — and arguably of more relevance to our present study — cannot be justified. Besides not being in accordance with how criminal investigators actually operate, if literally everyone in the entire world were genuinely a suspect, then justice could be served only by explicitly eliminating every single innocent person, an impossibly gargantuan task. Donnelly introduces the "*everyone* is a suspect" notion for one reason only, and it is not a legal reason. He introduces it because without it, his mathematics will not work.

Determination of a justifiable initial prior probability is, in fact, the single weak link in the Bayesian framework in many application of the method, not just criminal prosecution. But it is a major weakness, and is the reason why the majority of statisticians remain skeptical about the reliable applicability of the method, not just in legal cases but more generally. Once a starting value has been obtained, then, provided it is carried out correctly, with no "real world reasoning" getting mixed in (a difficult task to be sure, as was illustrated by the simple, toy example of the Monty Hall puzzle, but let us also leave that aside for now), Bayesian inference is a rock solid, 100% reliable method for assigning (consequent) probative value based on the additional evidence available.

Of course, the adage "garbage in, garbage out", familiar in the data processing world, is equally applicable in Bayesian reasoning. Unless the initial prior probability is an accurate representation of reality, the entire subsequent analysis is worthless.

But without a reliable starting value, the method never even gets off the ground.

In a cold hit case such as Jenkins, absent taking the RMP of the original identification search as starting value for *p* (and then of necessity not using it as Bayesian evidence in applying Bayes' theorem), the only way to carry through a Bayesian analysis is for someone to *assign* a starting value.

But who? How? On what basis? And under what legal authority?

As a mathematical logician, I find Bayesian reasoning internally consistent, coherent, and mathematically attractive.

Admittedly it is very hard to do correctly, but that should not prevent its use in courts of law, provided the payoff were better judgments.

However, I do not think the Bayesian approach can lead to better judgments, at least in Cold Hit DNA profile cases, and the reason is the framework's one major weakness, discussed above, concerning the determination of a starting value for the probability attached to the target proposition.

In the end, despite my enormous admiration for and appreciation of the mathematical elegance of the Bayesian methodology, and my recognition of the crucial role it can and does play in some domains — particularly of note these days the domain of Homeland Security — when it comes to the current debate about Cold Hit statistics, I find it hard not to see a strong similarity between the Bayesian camp and the ancient philosophers who devoted much time and intellectual fervor to a dialectic investigation, from basic principles of logic, into how many teeth a donkey has. According to the story, a young servant boy interrupted the learned men in their discussion and said, "Excuse me, gentlemen, I know you are all very clever, and much of what you say I don't understand, but why don't you just go outside, find a donkey, and count its teeth."

I believe that the story concludes with the philosophers, greatly angered by the young lad's impertinence, turning on him and running him out of town. So it is with some trepidation that I suggest that, in the case of deciding which statistics to use as evidence in a Cold Hit case, it's time to count the donkey's teeth. And that means adopting the frequentist-based approach of reasoning about the real world, not Bayesian reasoning about the evidence.

What can courts expect of juries in dealing with cold hit cases?

My personal interest in the question of how to assess the likelihood of a coincidental match in a Cold Hit DNA profiling search began in March 2005, when I was approached by the lawyers in the District of Columbia Public Defender Office who were representing Raymond Jenkins. I was initially somewhat surprised to be asked to help, since probability and statistics are not my particular areas of research specialization in mathematics. It turned out, however, that Jenkins' lawyers were interested in me because of my demonstrated abilities to take often difficult and arcane areas of advanced mathematics and make them intelligible to a lay audience, and my (associated) many years of experience (coming from written and verbal discussions with people who have read my various books and articles, attended talks by me, seen me on TV, or heard me on the radio) of the kinds of difficulties lay people typically encounter when faced with mathematical issues.

The court considering the Jenkins case had before it a considerable amount of background information and expert testimony, some of it contradictory, regarding the way the statisticians should compute the probability of error in a Cold Hit search. What the Jenkins' lawyers wanted from me was that I should interpret for them that body of information, and help them to understand it.

After examining the relevant materials, I wrote a brief affidavit to be submitted to the court, although at that point I had no idea what the case was that was under consideration.

My interest in the matter aroused, I subsequently undertook a deeper study of the issue. This paper is the result of that study. For the most part, the particular expertise I bring to the study is simply that of being a qualified, professional mathematician with 35 years of professional experience since I obtained my Ph.D. in mathematical logic in 1971. The one area where I do however bring what I believe is a fairly unique perspective and a notable level of expertise not shared by most other mathematicians is in my many years of familiarity with the kinds of difficulty that laypersons often have with mathematics — particularly questions of probability theory (a topic I have frequently written on and spoken about to lay audiences).

I believe that the perspective and experience I bring to this issue may be of value as the courts try to decide what statistical evidence to present to juries and how to assist them in weighing that evidence, but it does not and cannot carry the weight of a properly conducted scientific study of laypersons abilities to appreciate probabilistic figures. NRC II expressly called for such a study:

"Recommendation 6.1. Behavioral research should be carried out to identify any conditions that might cause a trier of fact to misinterpret evidence on DNA profiling and to assess how well various ways of presenting expert testimony on DNA can reduce any such misunderstandings."

NRC II's ensuing commentary on recommendation 6.1 reads, in part:

"… At present, policymakers must speculate about the ability of jurors to understand the significance of a match as a function of the method of presentation. Solid, empirical research into the extent to which the different methods advance juror understanding is needed."

Such a study is indeed urgently needed, in my view. Not least because the celebrated work of the psychologists Amos Tversky and Daniel Kahnemann in the 1970s and 80s illustrated just how poorly the average person appreciates and reasons with probabilities.[21]

[21] See Daniel Kahneman and Amos Tversky, "Prospect theory: An analysis of decisions under risk", *Econometrica*, 47:313327, 1979.

In my affidavit to the April 2005, *Jenkins* evidentiary hearing, I wrote:

> “Based on my 35 years of teaching mathematics, writing about it for experts and for laypersons, broadcasting about it, and receiving letters and emails from people from many walks of life, my belief is that many people, maybe even the majority, fundamentally DO NOT UNDERSTAND how probabilities work, even in the case of very simple examples. Moreover, in many cases, no amount of explanations by me or others can correct their false impressions. Well educated and otherwise rational people generally seem very certain about their probabilistic reasoning ability, even when their understanding is totally flawed, and they often exhibit a strong resistance to even acknowledging the possibility of their misunderstanding, let alone doing anything to correct it. (I have speculated in published writings elsewhere that there may be an evolutionary basis for this over-confidence.) Included among the many people who have confidence in an erroneous conception of probability are professors and teachers of mathematics and/or science. It denigrates no one for me to suggest that the chances of having even one person in a courtroom who really understands the issues surrounding probabilities, even in very simple cases, and even if experts are bought in to explain things, may be very low.
>
> The above remarks apply to even very simple uses of numerical probabilities. In the case of the use of the Cold Hit DNA matching technique, the issues are so subtle that even highly credentialed world experts can hold diametrically opposite views.”

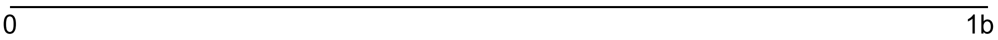
In my view, introducing figures such as “one in 26 quintillion” or “one in ten billion” or even “one in a million” into courtroom proceedings has almost no value, since no one, not even professional mathematicians, has any real sense of what such large numbers signify.

For example, do you know roughly how many seconds have elapsed since Christ was born?

Is that number greater or smaller than the distance from San Francisco to New York City measured in inches?

My point is not that anyone should know the answer to either question. Rather that most people simply have no idea *within several orders of magnitude*.

And why should they? We hardly ever have to deal with such large numbers, and never develop a sense of what they mean. For instance, below is a line representing one billion, with zero at the left, one billion at the right. Before reading any further, please make a pencil mark on the line where you think one million lies. There is no catch here. It’s not a trick question. Just put your mark where your intuition tells you it should go.



For the record, just over 63 billion seconds have elapsed since Christ was born, a number many, many orders of magnitude smaller than the 26 quintillion figure the prosecution is so adamant must be presented to the jury in the Jenkins trial.

The distance from San Francisco to New York City measured in inches is just under 163 million, much smaller than the number of seconds since Christ was alive.

As for that line, unless you placed your mark at the very left hand end of the line, you really don’t have any sense of a million and a billion and the difference between them. Measured along the line shown, running from 0 to 1 billion, a million is so close to 0 that it falls within the pencil mark you make at 0.

One final example to show just how ill equipped we are to handle large numbers and the processes that produce them. Imagine taking an ordinary 8×8 chess-board and placing piles of counters 2 mm thick (a Quarter is fairly close) on the squares according to the following rule. Number the squares from 1 to 64. On the first square place 2 counters. On square 2 place 4 counters. On square 3 place 8 counters. And so on, on each square placing exactly twice as many counters as on the previous one. How high do you think this pile on the final square will be? 1 meter? 100 meters? A kilometer?

As a matter of fact, your pile of counters will stretch out beyond the Moon (a mere 400,000 kilometers away) and the Sun (150 million kilometers away) and will in fact reach almost to the nearest star, Proxima Centauri, some 4 light years from Earth.

From a purely numerical point of view, a DNA profile match where the probability of it having come from someone other than the guilty individual is demonstrably less than 1 in ten million (say) *should be sufficient to convict*. Even the very best forensic laboratories could not guarantee their overall process to that degree of accuracy — no organization that depends on human performance can guarantee such precision. But given the possibility of confounding factors such as relatives (either known or unknown), in practice, it is surely reasonable and prudent to set the bar much higher, as is currently done.

But figures such as 1 in 26 quintillion are surely massive overkill. Why then are prosecuting attorneys so eager to quote such ridiculously large numbers in making their case that an accused is guilty? One can only assume that it is for rhetorical effect; they hope that by presenting the jury with a number so large it defies all imagination, the jury members — the majority of whom are likely not only to have limited mathematical skills but in fact be intimidated by the subject — will be sufficiently overwhelmed that they will not question the evidence, and how it was arrived at.

But such use of numbers and mathematics — a system that when properly used provides humanity with an extraordinary degree of precision in many walks of life — makes a mockery of science and mathematics and is both intellectually dishonest and morally indefensible.

Pending the results of the kind of study that NRC II advocates in Recommendation 6.1 (which I anticipate will be such that what I am about to say will remain valid), all my experience in dealing with laypersons in matters of probabilistic reasoning leads me to suggest that the evidentiary introduction of probabilities in court cases is *always* as likely to be incorrectly interpreted as correctly, and thus should be kept out as far as is reasonably possible. Both because many people experience enormous difficulty understanding probability, even in simple cases and even when others try to help them, and because the potential for the misuse of probabilities by rhetorically skillful lawyers is simply too great.

Far better simply to present jurors with evidence that is primarily qualitative, with any numbers mentioned being *within the comprehension of the average person*. For example, “Members of the jury, the DNA match indicates that X is the guilty person. Like any human process, the process whereby the DNA match was obtained carries with it a small chance of error. However, a body of experts appointed by the National Academy of Sciences has examined the procedures used, and they testify that the likelihood of an error here is so small that if you were to serve on a jury considering such a case every day for a million years, you would almost certainly never encounter a case where the DNA evidence identified the wrong person.”

Demystifying the Arizona database study

To drive home the point about how laypersons are likely to misinterpret the large figures often bandied around in courts, let me provide the explanation I promised earlier as to why the results of the Arizona database survey (see pages 9, 14, and 38), though seeming to contract the mathematics, are actually exactly what the mathematics predicts.

I suspect many laypeople are likely to reason as follows: “Since DNA profiling has an inaccuracy rate less than 1 in many trillion (or more), the chances of there being a false match in a database of maybe 3 million entries, such as the CODIS database, is so tiny that no matter which method you use to calculate the odds, a match will surely be definitive proof." The intuition behind such a conclusion is presumably that the database search has 3 million shots at finding a match, so if the odds against there being a match are 1 in 10 trillion, then the odds against finding a match in the entire database are roughly 1 in 3 million (3 million divided by 3 trillion is roughly 1/3,000,000).

Unfortunately — at least it could be for an innocent defendant in the case — this argument is not valid. In fact, notwithstanding an RMP in the "1 in many trillion" range, even a fairly small DNA database is likely to contain many pairs of accidental matches, where two different people have the same DNA profile. A tiny RMP simply does not mean there won't be accidental matches. The argument is the same as the one used in the famous Birthday Paradox (actually not a paradox, just a surprise) that you need only 23 randomly selected people in a room in order for the probability that two of them share a birthday to be greater than one-half. (The probably works out to be 0.507.)

The Arizona DNA convicted offender database has some 65,000 entries, each entry being a 13 loci profile. Suppose, for simplicity, that the probability of a random match at a single locus is 1/10, a figure that, as we observed earlier, is not unreasonable. Thus the RMP for a 9 locus match is $1/10^9$, i.e., 1 in 1 billion. You might think that with such long odds against a randomly selected pair of profiles matching at 9 loci, it would be highly unlikely that the database contained a pair of entries that were identical on 9 loci. Yet, by the same argument used in the Birthday Puzzle, the probability of getting two profiles that match on 9 loci is around 5%, or 1 in 20. For a database of 65,000 entries, that means you would be quite likely to find some matching profiles!

Before I sketch the calculation, I'll note that the answer becomes less surprising when you realize that for a database of 65,000 entries, there are roughly $65,000^2$ = 4,225,000,000 (just over 4 billion) possible pairs of entries, each one of which has a chance of yielding a 9-loci match.

Because the database calculation involves very large numbers, I'll first of all go through the math that resolves the Birthday Paradox itself.

The question, remember, is how many people you need to have at a party so that there is a better-than-even chance that two of them will share the same birthday? Most people think the answer is 183, the smallest whole number larger than 365/2. The number 183 is the correct answer to a very different question: How many people do you need to have at a party so that there is a better-than-even chance that one of them will share ***your*** birthday? If there is no restriction on which two people will share a birthday, it makes an enormous difference. With 23 people in a room, there are 253 different ways of pairing two people together, and that gives a lot of possibilities of finding a pair with the same birthday.

To figure out the exact probability of finding two people with the same birthday in a given group, it turns out to be easier to ask the opposite question: what is the probability that NO two will share a birthday, i.e., that they will all have different birthdays? With just two people, the probability that they have different birthdays is 364/365, or about .997. If a third person joins them, the probability that this new person has a different birthday from those two (i.e., the probability that all three will have different birthdays) is (364/365) x (363/365), about .992. With a fourth person, the probability that all four have different birthdays is (364/365) x (363/365) x (362/365), which comes out at around .983. And so on. The answers to these multiplications get steadily smaller. When a twenty-third person enters the room, the final fraction that you multiply by is 343/365, and the answer you get drops below .5 for the first time, being approximately .493. This is the probability that all 23 people have a different birthday. So, the probability that at least two people share a birthday is 1 - .493 = .507, just greater than 1/2.

The following table shows what happens for some other values of n, the number of randomly selected people in the room.

n	Probability of at least one match
23	50.7%
25	56.9%
30	70.6%
35	81.4%
40	89.1%
45	94.1%
50	97.0%

You need only have 50 people to make getting a coincidence an odd-on certainty!

Now for the Arizona database. The reason you can expect 9-locus matches is the same as for the coincident birthdays, but the numbers involved are much bigger, and accordingly I'll present the calculation in a slightly different way.

Recall that we have a DNA profile database with 65,000 entries, each entry being a 13-loci profile. We suppose that the probability of a random match at a single locus is 1/10, so the RMP for a 9 locus match is $1/10^9$, i.e., 1 in billion.

Now, there are 13!/[9! x 4!] = [13 x 12 x 11 x 10]/[4 x 3 x 2 x 1] = 715 possible ways to choose 9 loci from 13, so the RMP for finding a match on *any* 9 loci of the 13 is $715/10^9$.

If you pick any profile in the database, the probability of a second profile not matching on 9 loci is roughly $1 – 715/10^9$.

Hence, the probability of all 65,000 database entries not matching on 9 loci is roughly $(1 – 715/10^9)^{65,000}$. Using the binomial theorem, this is approximately 1 – 65,000 x $715/10^9$ = 1 – $46,475/10^6$, roughly 1 – .05.

The probability of there being a 9-locus match is the difference between 1 and this figure, namely 1 – (1 – 0.05) = 0.05. That's roughly a 5% chance.

So the results found in the Arizona database study should not come as a surprise to anyone who understands the mathematics. But can we *really* expect the average judge and juror to follow the above calculation? Some may argue yes, but my thirty years teaching mathematics *at university level*, never mind my experience explaining math to wider, lay audiences tells me that such an assumption is totally unrealistic.

APPENDIX Bayes' theorem and Bayesian inference

Bayesian analysis depends on a mathematical theorem proved by an 18th Century English Presbyterian minister by the name of Thomas Bayes. Bayes' theorem languished largely ignored and unused for over two centuries (in large part because of its dependence on an initial prior probability figure, for which there is often no justifiable means of determination) before statisticians, lawyers, medical researchers, software developers, and others started to use it in earnest during the 1990s.

What makes this relatively new technique of "Bayesian inference" particularly intriguing is that it uses an honest-to-goodness mathematical formula (Bayes' Theorem) in order to improve — on the basis of evidence — the best (human) estimate that a given proposition is true. In the words of some statisticians, it's "mathematics on top of common sense." You start with an initial estimate of the probability that the proposition is true and an estimate of the reliability of the evidence. The method then tells you how to combine those two figures — in a precise, mathematical way — to give a new estimate of the probability the proposition is true in the light of the evidence.

In some highly constrained situations, both initial estimates may be entirely accurate, and in such cases Bayes' method will give you the correct answer. (For example, in resolving the Monty Hall puzzle and its variants.)

In a more typical real-life situation, you don't have exact figures, but as long as the initial estimates are reasonably good, then the method will give you a better estimate of the probability that the event of interest will occur. Thus, in the hands of an expert in the domain under consideration, someone who is able to assess all the available evidence reliably, Bayes' method can be a powerful tool.

In general, Bayes' method shows you how to calculate the probability (or improve an estimate of the probability) that a certain proposition S is true, based on evidence for S, when you know (or can estimate):

(1) the probability of S in the absence of any evidence;

(2) the evidence for S;

(3) the probability that the evidence would arise regardless of whether S or not;

(4) the probability that the evidence would arise if S were true.

The key formula that tells you how to update your probability (estimate) is given by Bayes' theorem, which I outline below.

Let P(S) be the numerical probability that the proposition S is true in the absence of any evidence. P(S) is known as the *prior probability*.

You obtain some evidence, E, for S.

Let P(S|E) be the probability that S is true given the evidence E. This is the revised estimate you want to calculate.

A quantity such as P(S|E) is known as a *conditional probability* — the conditional probability of S being true, given the evidence E.

Let P(E) be the probability that the evidence E would arise if S were not known to be true and let P(E|S) be the probability that E would arise if S were true.

The ratio P(E|S)/P(E) is called the *likelihood ratio* for E given S.

Bayes' theorem says that the posterior probability P(S|E) is derived from the prior probability P(S) by multiplying the latter by the likelihood ratio for E given S:

$$P(S|E) = P(S) \times P(E|S) / P(E)$$

Notice how the formula reduces the problem of computing how probable S is, given the evidence, to computing how probable it would be that the evidence arises if S were true.

To indicate just *how* difficult it can be to carry our correct Bayesian reasoning, using Bayes' theorem, you might like to try using the above formula in order to resolve the two variants of the Monty Hall problem, the original version where Monty always opens an unchosen door that has no prize (where the answer is that the probability of winning doubles from 1/3 to 2/3 if you switch) and the variant where another contestant opens an unchosen door (and the answer is that it makes no difference whether you switch or stick).[22]

In any court case where conditional probabilities are introduced, juries need to be careful not to confuse the very different probabilities

- P(G|E), the conditional probability that the defendant is guilty given the evidence;
- P(E|G), the conditional probability that the evidence would be found assuming the defendant were guilty;
- P(E), the probability that the evidence could be found among the general population

The figure of relevance in deciding guilt is P(G|E).

As Bayes' formula shows, P(G|E) and P(E) can be very different, with P(G|E) generally much lower than P(E). This is why the introduction of P(E) in court proceedings is so undesirable.

Bayesian inference methods lie behind a number of new products on the market. For example, chemists can take advantage of a software system that uses Bayesian methods to improve the resolution of nuclear magnetic resonance (NMR) spectrum data. Chemists use such data to work out the molecular structure of substances they wish to analyze. The system uses Bayes' formula to combine the new data from the NMR device with existing NMR data, a procedure that can improve the resolution of the data by several orders of magnitude.

[22] In the variant game, it is not the case that the opening of an unchosen door by the second contestant does not change what you know. If the door opened reveals the prize, then of course your knowledge has changed; and what is more, the game is over. But even if that opened door does not hide the prize, its opening changes the posterior probabilities. However, whereas the posterior probabilities that the prize is behind door A or door B after Monty opens door C in the standard Monty Hall game are 1/3 and 2/3, respectively, making it wise to switch, in the variant, if the other contestant opens door C and there is no prize, the posterior probabilities are 1/2 for door A and 1/2 for door B (the prior probabilities for all three doors are 1/3, of course), and thus it does not matter whether you switch or stick.

Other recent uses of Bayesian inference are in the evaluation of new drugs and medical treatments and the analysis of police arrest data to see if any officers have been targeting one particular ethnic group.

Bibliography

NRC I: *DNA Technology in Forensic Science*, National Academy Press, 1992,

NRC II: *The Evaluation of Forensic DNA Evidence*, National Academy Press, 1996

David Balding and Peter Donnelly, “Inferring Identity from DNA Profile Evidence”, *Proc. Natl. Acad. Sci. USA*, Vol 92, pp.11741–11745, 1995.

David Balding and Peter Donnelly, “Evaluating DNA Profile Evidence When the Suspect is Identified Through a Database Search”, *J. Forensic Science* 603, 1996

Peter Donnelly, “DNA Evidence after a database hit”, affidavit submitted to the Superior Court of the District of Columbia, October 3, 2004.

Peter Donnelly and Richard Friedman, “DNA Database Searches And The Legal Consumption Of Scientific Evidence”, *Michigan Law Review*, 00262234, Vol. 97, Issue 4, February 1999.

DNA Advisory Board, “Statistical and Population Genetics Issues Affecting the Evaluation of the Frequency of Occurrence of DNA Profiles Calculated From Pertinent Population Database(s)”, *Forensic Science Communications*, July 2000, Volume 2, Number 3, U.S. Department of Justice, FBI.

Daniel Kahneman and Amos Tversky, “Prospect theory: An analysis of decisions under risk”, *Econometrica*, 47:313327, 1979.

Public Defender Service, District of Columbia, *United States of America v. Raymond Jenkins*, “Reply To Government’s Opposition To Motion To Exclude DNA “Inclusion” Evidence, Expert Testimony, And Frequency Statistics And Government’s Opposition To Supplemental Motion To Exclude Nuclear DNA “Match” Evidence, DNA Expert Testimony, And “Random Match Probability” “ brief submitted to the Superior Court of the District of Columbia, October 12, 2004.

Public Defender Service, District of Columbia, *United States of America v. Raymond Jenkins*, “Appeal from the Superior Court of the District of Columbia”, brief submitted to the District of Columbia Court of Appeals, July 28, 2005.