Module 7: Probabilistic (Mis)Reasoning and Related Confusions

bioinformatics: a synergistic fusion of huge data bases and bad statistics
data mining: panning for gold in a sewer
– Stephen Senn (*Dicing with Death*, 2003)

Abstract: The introductory module started with the well-known case of Sally Clark and how a misunderstanding about probabilistic independence helped lead to her wrongful imprisonment for killing her two children. The present module will provide more examples of mistaken probabilistic reasoning, with many involving misinterpretations of conditional probability. We will revisit the O.J. Simpson criminal case where his defense team took advantage of what is termed the "defendant's fallacy," as well as some specious reasoning about conditional probability (perpetrated by Alan Dershowitz). Several additional high-profile legal cases will be mentioned that were mishandled because of the prosecutor's fallacy, much like that of Sally Clark. One is recent – the Dutch nurse, Lucia de Berk, was accused of multiple deaths at the hospitals she worked at in the Netherlands; another is much older and involves the turn-of-the-century (the late 1800s, that is) case of Alfred Dreyfus, the much maligned French Jew who was falsely imprisoned for espionage.

Contents

1 The (Mis)assignment of Probabilities

Clear probabilistic reasoning requires a good understanding of how conditional probabilities are defined and operate. There are many day-to-day contexts we face where decisions might best be made from conditional probabilities, if we knew them, instead of from marginal information. When deciding on a particular medical course of action, for example, it is important to condition on personal circumstances of age, risk factors, family medical history, and our own psychological needs and makeup. A fairly recent and controversial instance of this, where the conditioning information is "age," is reported in the *New York Times* article by Gina Kolata, "Panel Urges Mammograms at 50, Not 40" (November 16, 2009). The failure to consider conditional instead of marginal probabilities is particularly grating for many of us who follow various sporting activities and enjoy second-guessing managers, quarterbacks, sports commentators, and their ilk. As an example, consider the "strike-'em-out-throw-'em-out" double play in baseball, where immediately after the batter has swung and missed at a third strike or taken a called third strike, the catcher throws out a base runner attempting to steal second or third base. Before such a play occurs, announcers routinely state that the runner "will or will not be sent" because the "batter strikes out only some percentage of the time." The issue of running or not shouldn't be based on the marginal probability of the batter striking out but on some conditional probability (for example, how often does the batter strike out

when faced with a particular count or type of pitcher). For many other instances, however, we might be content not to base our decisions on conditional information; for example, always wear a seat belt irrespective of the type or length of trip being taken.

Although the assignment of probabilities to events consistent with the mutually exclusive event rule may lead to an internally valid system mathematically, there is still no assurance that this assignment is "meaningful," or bears any empirical validity for observable long-run expected frequencies. There seems to be a never-ending string of misunderstandings in the way probabilities can be generated that are either blatantly wrong, or more subtly incorrect, irrespective of the internally consistent system they might lead to. Some of these problems are briefly sketched below, but we can only hope to be representative of a few possibilities, not exhaustive.

One inappropriate way of generating probabilities is to compute the likelihood of some joint occurrence after some of the outcomes are already known. For example, there is the story about the statistician who takes a bomb aboard a plane, reasoning that if the probability of one bomb on board is small, the probability of two is infinitesimal. Or, during World War I, soldiers were actively encouraged to use fresh shell holes as shelter because it was very unlikely for two shells to hit the same spot during the same day. And the Minnesota Twins baseball manager who bats for an individual who earlier in the game hit a home run because it would be very unlikely for him to hit two home runs in the same game. Although these slightly amusing stories may provide obvious misassignments of probabilities, other related situations are more subtle. For example, whenever coincidences are

culled or "hot spots" identified from a search of available information, the probabilities that are then regenerated for these situations may not be valid. There are several ways of saying this: when some set of observations is the source of an initial suspicion, those same observations should not be used in a calculation that then tests the validity of the suspicion. In Bayesian terms, you should not obtain the posterior probabilities from the same information that gave you the prior probabilities. Alternatively said, it makes no sense to do formal hypothesis assessment by finding estimated probabilities when the data themselves have suggested the hypothesis in the first place. Some cross-validation strategy is necessary; for example, collecting independent data. Generally, when some process of search or optimization has been used to identify an unusual situation (for instance, when a "good" regression equation is found through a step-wise procedure [see Freedman, 1983, for a devastating critique]; when data are "mined" and unusual patterns identified; when DNA databases are searched for "cold-hits" against evidence left at a crime scene; when geographic "hot spots" are identified for, say, some particularly unusual cancer; or when the whole human genome is searched for clues to common diseases), the same methods for assigning probabilities before the particular situation was identified are generally no longer appropriate after the fact.[1]

A second general area of inappropriate probability assessment con-

[1]A particularly problematic case of culling or locating "hot spots" is that of residential cancer-cluster identification. A readable account is by Atul Gawande, "The Cancer-Cluster Myth," *New Yorker*, February 8, 1999. For the probability issues that arise in searching the whole human genome for clues to some condition, see "Nabbing Suspicious SNPS: Scientists Search the Whole Genome for Clues to Common Diseases" (Regina Nuzzo, *ScienceNews*, June 21, 2008).

cerns the model postulated to aggregate probabilities over several events. Campbell (1974, p. 126) cites an article in the *New York Herald Tribune* (May, 1954) stating that if the probability of knocking down an attacking airplane were .15 at each of five defensive positions before reaching the target, then the probability of knocking down the plane before it passed all five barriers would be .75 ($5 \times .15$), this last value being the simple sum of the individual probabilities—and an inappropriate model. If we could correctly assume independence between the Bernoulli trials at each of the five positions, a more justifiable value would be one minus the probability of passing all barriers successfully: $1.0 - (.85)^5 \approx .56$. The use of similar binomial modeling possibilities, however, may be specious—for example, when dichotomous events occur simultaneously in groups (such as in the World Trade Center disaster on 9/11/01); when the success proportions are not valid; when the success proportions change in value over the course of the trials; or when time dependencies are present in the trials (such as in tracking observations above and below a median over time). In general, when wrong models are used to generate probabilities, the resulting values may have little to do with empirical reality. For instance, in throwing dice and counting the sum of spots that result, it is not true that each of the integers from two through twelve is equally likely. The model of what is equally likely may be reasonable at a different level (for example, pairs of integers appearing on the two dice), but not at all aggregated levels. There are some stories, probably apocryphal, of methodologists meeting their demises by making these mistakes for their gambling patrons.

Flawed calculations of probability can have dire consequences within

our legal systems, as the case of Sally Clark and related others make clear. One broad and current area of possible misunderstanding of probabilities is in the context of DNA evidence (which is exacerbated in the older and more fallible system of identification through fingerprints).[2] In the use of DNA evidence (and with fingerprints), one must be concerned with the Random Match Probability (RMP): the likelihood that a randomly selected unrelated person from the population would match a given DNA profile. Again, the use of independence in RMP estimation is questionable; also, how does the RMP relate to, and is it relevant for, "cold-hit" searches in DNA databases. In a confirmatory identification case, a suspect is first identified by non-DNA evidence; DNA evidence is then used to corroborate traditional police investigation. In a "cold-hit" framework, the suspect is first identified by a search of DNA databases; the DNA evidence is thus used to identify the suspect as perpetrator, to the exclusion of others, directly from the outset (this is akin to shooting an arrow into a tree and then drawing a target around it). Here, traditional police work is no longer the focus. For a thorough discussion of the probabilistic context surrounding DNA evidence, which extends with even greater force to fingerprints, the article by Jonathan Koehler is recommended ("Error and Exaggeration in the Presentation of DNA Evidence at Trial," *Jurimetrics Journal*, *34*, 1993–1994, 21–39). We excerpt part of the introduction to this article below:

DNA identification evidence has been and will continue to be powerful evidence against criminal defendants. This is as it should be. In general, when blood, semen or hair that reportedly matches that of a defendant is found on

[2]Two informative articles on identification error using fingerprints ("Do Fingerprints Lie?", Michael Specter, *New Yorker*, May 27, 2002), and DNA ("You Think DNA Evidence is Foolproof? Try Again," Adam Liptak, *New York Times*, March 16, 2003).

or about a victim of violent crime, one's belief that the defendant committed the crime should increase, based on the following chain of reasoning:

Match Report ⇒ True Match ⇒ Source ⇒ Perpetrator

First a reported match is highly suggestive of a true match, although the two are not the same. Errors in the DNA typing process may occur, leading to a false match report. Second, a true DNA match usually provides strong evidence that the suspect who matches is indeed the source of the trace, although the match may be coincidental. Finally, a suspect who actually is the source of the trace may not be the perpetrator of the crime. The suspect may have left the trace innocently either before or after the crime was committed.

In general, the concerns that arise at each phase of the chain of inferences are cumulative. Thus, the degree of confidence one has that a suspect is the source of a recovered trace following a match report should be somewhat less than one's confidence that the reported match is a true match. Likewise, one's confidence that a suspect is the perpetrator of a crime should be less than one's confidence that the suspect is the source of the trace.

Unfortunately, many experts and attorneys not only fail to see the cumulative nature of the problems that can occur when moving along the inferential chain, but they frequently confuse the probabilistic estimates that are reached at one stage with estimates of the others. In many cases, the resulting misrepresentations and misinterpretation of these estimates lead to exaggerated expressions about the strength and implications of the DNA evidence. These exaggerations may have a significant impact on verdicts, possibly leading to convictions where acquittals might have been obtained.

This Article identifies some of the subtle, but common, exaggerations that have occurred at trial, and classifies each in relation to the three questions that are suggested by the chain of reasoning sketched above: (1) Is a reported match a true match? (2) Is the suspect the source of the trace? (3) Is the suspect the perpetrator of the crime? Part I addresses the first question and discusses ways of defining and estimating the false positive error rates at DNA laboratories. Parts II and III address the second and third questions, respectively. These sections introduce the "source probability error" and "ultimate issue error" and show how experts often commit these errors at

trial with assistance from attorneys on *both* sides. (pp. 21–22)

In 1989, and based on urging from the FBI, the National Research Council (NRC) formed the Committee on DNA Technology in Forensic Science, which issued its report in 1992 (*DNA Technology in Forensic Science*; or more briefly, NRC I). The NRC I recommendation about the cold-hit process was as follows:

> 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). (p. 124)

A follow-up report by a second NRC panel was published in 1996 (*The Evaluation of Forensic DNA Evidence*; or more briefly, NRC II), having the following main recommendation about cold-hit probabilities and using the "database match probability" or DMP:

> 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. (p. 161)

The term "database match probability" (DMP) is somewhat unfortunate. This is not a real probability but more of an expected number of matches given the RMP. A more legitimate value for the probability that another person matches the defendant's DNA profile would be $1 - (1 - \frac{1}{\text{RMP}})^N$, for a database of size N; that is, one minus the probability of no matches over N trials. For example, for an RMP of 1/1,000,000 and an N of 1,000,000, the above probability of another

match is .632; the DMP (not a probability) number is 1.00, being the product of N and RMP. In any case, NRC II made the recommendation of using the DMP to give a measure of the accuracy of a cold-hit match, and did not support the more legitimate "probability of another match" using the formula given above (possibly because it was considered too difficult?):[3]

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)

2 More on Bayes' Rule and the Confusion of Conditional Probabilities

The case of Sally Clark discussed in the introductory module and the commission of the prosecutor's fallacy that lead to her conviction is not an isolated occurrence. There was the recent miscarriage of justice in the Netherlands involving a nurse, Lucia de Berk, accused of

[3]As noted repeatedly by Gigerenzer and colleagues (e.g., Gigerenzer, 2002; Gigerenzer et al., 2007), it also may be best for purposes of clarity and understanding, to report probabilities using "natural frequencies." For example, instead of saying that a random match probability is .01, this could be restated alternatively that for this population, 1 out of every 10,000 men would be expected to show a match. The use of natural frequencies supposedly provides a concrete reference class for a given probability that then helps interpretation.

multiple deaths at the hospitals where she worked. This case aroused the international community of statisticians to redress the apparent injustices visited upon Lucia de Berk. One source for background, although now somewhat dated, is Mark Buchanan at the *New York Times* online opinion pages ("The Prosecutor's Fallacy," May 16, 2007). The Wikipedia article on Lucia de Berk provides the details of the case and the attendant probabilistic arguments, up to her complete exoneration in April 2010.

A much earlier and historically important *fin de siecle* case, is that of Alfred Dreyfus, the much maligned French Jew, and captain in the military, who was falsely imprisoned for espionage. In this case, the nefarious statistician was Alphonse Bertillon, who through a very convoluted argument reported a small probability that Dreyfus was "innocent." This meretricious probability had no justifiable mathematical basis and was generated from culling coincidences involving a document, the handwritten *bordereau* (without signature) announcing the transmission of French military information. Dreyfus was accused and convicted of penning this document and passing it to the (German) enemy. The "prosecutor's fallacy" was more or less invoked to ensure a conviction based on the fallacious small probability given by Bertillon. In addition to Émile Zola's well-known article, *J'accuse* ... !, in the newspaper *L'Aurore* on January 13, 1898, it is interesting to note that turn-of-the-century well-known statisticians and probabilists from the French Academy of Sciences (among them Henri Poincaré) demolished Bertillon's probabilistic arguments, and insisted that any use of such evidence needs to proceed in a fully Bayesian manner, much like our present understanding of evidence in current forensic science and the proper place of probabilistic argu-

mentation.[4]

We observe the same general pattern in all of the miscarriages

[4]By all accounts, Bertillon was a dislikable person. He is best known for the development of the first workable system of identification through body measurements; he named this "anthropometry" (later called "bertillonage" by others). We give a brief quotation about Bertillon from *The Science of Sherlock Holmes* by E. J. Wagner (2006):

And then, in 1882, it all changed, thanks to a twenty-six-year old neurasthenic clerk in the Paris Police named Alphonse Bertillon. It is possible that Bertillon possessed some social graces, but if so, he was amazingly discreet about them. He rarely spoke, and when he did, his voice held no expression. He was bad-tempered and avoided people. He suffered from an intricate variety of digestive complaints, constant headaches, and frequent nosebleeds. He was narrow-minded and obsessive.

Although he was the son of the famous physician and anthropologist Louis Adolphe Bertillon and had been raised in a highly intellectual atmosphere appreciative of science, he had managed to be thrown out of a number of excellent schools for poor grades. He had been unable to keep a job. His employment at the police department was due entirely to his father's influence. But this misanthropic soul managed to accomplish what no one else had: he invented a workable system of identification.

Sherlock Holmes remarks in *The Hound of the Baskervilles*, "The world is full of obvious things which nobody by any chance ever observes." It was Bertillon who first observed the obvious need for a scientific method of identifying criminals. He recalled discussions in his father's house about the theory of the Belgian statistician Lambert Adolphe Jacques Quetelet, who in 1840 had suggested that there were no two people in the world who were exactly the same size in all their measurements. (pp. 97–98)

Bertillonage was widely used for criminal identification in the decades surrounding the turn-of-the-century. It was eventually supplanted by the use of fingerprints, as advocated by Sir Francis Galton in his book, *Finger Prints*, published in 1892. A short extraction from Galton's introduction mentions Bertillon by name:

My attention was first drawn to the ridges in 1888 when preparing a lecture on Personal Identification for the Royal Institution, which had for its principal object an account of the anthropometric method of Bertillon, then newly introduced into the prison administration of France. Wishing to treat the subject generally, and having a vague knowledge of the value sometimes assigned to finger marks, I made inquiries, and was surprised to find, both how much had been done, and how much there remained to do, before establishing their theoretical value and practical utility.

One of the better known photographs of Galton (at age 73) is a Bertillon record from a visit Galton made to Bertillon's laboratory in 1893 (a Google search using the two words "Galton" and "Bertillon" will give the image).

of justice involving the prosecutor's fallacy. A very small reported probability of "innocence" is reported, typically obtained incorrectly either by culling, misapplying the notion of statistical independence, or using an inappropriate statistical model. This probability is calculated by a supposed expert with some credibility in court: Roy Meadow for Clark, Henk Elffers for de Berk, Alphonse Bertillon for Dreyfus. The prosecutor's fallacy then takes place, leading to a conviction for the crime. Various outrages ensue from the statistically literate community, with the eventual emergence of some "statistical good guys" hoping to redress the wrongs done: Richard Gill for de Berk, Henri Poincaré (among others) for Dreyfus, the Royal Statistical Society for Clark. After long periods of time, convictions are eventually overturned, typically after extensive prison sentences have already been served. We can only hope to avoid similar miscarriages of justice in cases yet to come by recognizing the tell-tale pattern of occurrences for the prosecutor's fallacy.

Any number of conditional probability confusions can arise in important contexts and possibly when least expected. A famous instance of such a confusion was in the O.J. Simpson case, where one conditional probability, say, $P(A|B)$, was equated with another, $P(A|B \text{ and } D)$. We quote the clear explanation of this obfuscation

Besides anthropometry, Bertillon contributed several other advances to what would now be referred to as "forensic science." He standardized the criminal "mug shot," and the criminal evidence picture through "metric photography." Metric photography involves taking pictures before a crime scene is disturbed; the photographs had mats printed with metric frames placed on the sides. As in "mug shots," photographs are generally taken of both the front and side views of a scene. Bertillon also created other forensic techniques, for example, forensic document examination (but in the case of Dreyfus, this did not lead to anything good), the use of galvanoplastic compounds to preserve footprints, the study of ballistics, and the dynamometer for determining the degree of force used in breaking and entering.

by Krämer and Gigerenzer (2005):

> Here is a more recent example from the U.S., where likewise $P(A|B)$ is confused with $P(A|B \text{ and } D)$. This time the confusion is spread by Alan Dershowitz, a renowned Harvard Law professor who advised the O.J. Simpson defense team. The prosecution had argued that Simpson's history of spousal abuse reflected a motive to kill, advancing the premise that "a slap is a prelude to homicide." Dershowitz, however, called this argument "a show of weakness" and said: "We knew that we could prove, if we had to, that an infinitesimal percentage—certainly fewer than 1 of 2,500—of men who slap or beat their domestic partners go on to murder them." Thus, he argued that the probability of the event K that a husband killed his wife if he battered her was small, $P(K|\text{battered}) = 1/2{,}500$. The relevant probability, however, is not this one, as Dershowitz would have us believe. Instead, the relevant probability is that of a man murdering his partner given that he battered her and that she was murdered, $P(K|\text{battered and murdered})$. This probability is about 8/9. It must of course not be confused with the probability that O.J. Simpson is guilty; a jury must take into account much more evidence than battering. But it shows that battering is a fairly good predictor of guilt for murder, contrary to Dershowitz's assertions. (p. 228)

Avoiding the prosecutor's fallacy is one obvious characteristic of correct probabilistic reasoning in legal proceedings. A related specious argument on the part of the defense is the "defendant's fallacy" (Committee on DNA Technology in Forensic Science, 1992, p. 31). Suppose that for an accused individual who is innocent, there is a one-in-a-million chance of a match (such as for DNA, blood, or fiber). In an area of, say, 10 million people, the number of matches expected is 10 even if everyone tested is innocent. The defendant's fallacy would be to say that because 10 matches are expected in a city of 10 million, the probability that the accused is innocent is 9/10. Because this latter probability is so high, the evidence of a match for

the accused cannot be used to indicate a finding of guilt, and therefore, the evidence of a match should be excluded. A version of this fallacy appeared (yet again) in the O.J. Simpson murder trial; we give a short excerpt about the defendant's fallacy that is embedded in the Wikipedia article on the prosecutor's fallacy :

A version of this fallacy arose in the context of the O.J. Simpson murder trial where the prosecution gave evidence that blood from the crime scene matched Simpson with characteristics shared by 1 in 400 people. The defense retorted that a football stadium could be filled full of people from Los Angeles who also fit the grouping characteristics of the blood sample, and therefore the evidence was useless. The first part of the defenses' argument that there are several other people that fit the blood grouping's characteristics is true, but what is important is that few of those people were related to the case, and even fewer had any motivation for committing the crime. Therefore, the defenses' claim that the evidence is useless is untrue.

We end this chapter with two additional fallacies involving conditional probabilities that were also reviewed by Krämer and Gigerenzer (2005). One will be called the facilitation fallacy, and the second, the category (mis)representation fallacy.

The facilitation fallacy argues that because a conditional probability, $P(B|A)$, is "large," the event B must therefore be facilitative for A (i.e., it must be true that $P(A|B) > P(A)$). As an example, suppose that among all people involved in an automobile accident, the majority are male; or, $P(\text{male}|\text{accident})$ is "large." But this does not imply that being male is facilitative of having an accident (i.e., it is not necessarily true that $P(\text{accident}|\text{male}) > P(\text{accident})$. There could be, for example, many more male drivers on the road than female drivers, and even though accident rates per mile may be the

same for males and females, males will be in the majority when only those individuals involved in an accident are considered.

The category (mis)representation fallacy begins with the true observation that if B is facilitative of A, so that $P(A|B) > P(A)$, then $\bar{B}$ must be inhibitive of A; that is, $P(A|\bar{B}) < P(A)$. The fallacy is to then say that all subsets of $\bar{B}$ must also be inhibitive of A as well.

To paraphrase a hypothetical example given by Krämer and Gigerenzer(2005), suppose an employer hires 158 out of 1000 applicants (among the 1000, 200 are black, 200 are Hispanic, and 600 are white). Of the 158 new hires, 38 are non-white (36 are Hispanic and 2 are black), and 120 are white. Being white is facilitative of being hired:

$P(\text{hired}|\text{white}) = \frac{120}{600} = .20 > P(\text{hired}) = \frac{158}{1000} = .158$

And being nonwhite is inhibitive of being hired:

$P(\text{hired}|\text{nonwhite}) = \frac{38}{400} = .095 < P(\text{hired}) = .158$

But note that although being black is inhibitive of being hired:

$P(\text{hired}|\text{black}) = \frac{2}{200} = .01 < P(\text{hired}) = .158,$

the same is not true for the Hispanic subset:

$P(\text{hired}|\text{Hispanic}) = \frac{36}{200} = .18$ is greater than $P(\text{hired}) = .158.$

References

[1] Campbell, S. K. (1974). *Flaws and fallacies in statistical thinking*. Englewood Cliffs, NJ: Prentice-Hall.

[2] Freedman, D. A. (1983). A note on screening regression equations. *American Statistician*, *37*, 152–155.

[3] Gigerenzer, G. (2002). *Calculated risks: How to know when numbers deceive you*. New York: Simon & Schuster.

[4] Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, *8*, 53–96.

[5] Krämer, W., & Gigerenzer, G. (2005). How to confuse with statistics or: The use and misuse of conditional probabilities. *Statistical Science*, *20*, 223–230.