2011

Statistics In Law: Bad Inferences & Uncommon Sense

Curtis E.A. Karnow

Statistics In Law: Bad Inferences & Uncommon Sense

By

Curtis E.A. Karnow[π]

"We do not convict people in these courts on statistics. It would be a terrible day if that were so."
Mr. Justice Harrison, Presiding Judge at trial at which Sally Clark was convicted based on misuse of statistical evidence

"Common sense is the collection of prejudices acquired by age eighteen."
Albert Einstein

I. *Opening Sequence*

1. In 1995, O.J. Simpson stood trial for the murder of his wife. One of his attorneys Alan Dershowitz reported—correctly—that fewer than one in 1000 wives who are abused by their husbands are later killed by the spouse. Dershowitz, who teaches criminal law at Harvard Law School, argued this made it entirely unlikely his client had committed the murder of Simpson's admittedly battered wife. While there were probably many reasons for the result, Simpson was acquitted.

2. In 1999 Sally Clark was convicted of murdering her two sons. One died at eight weeks, another at eleven weeks, both initially thought to have died of sudden infant death syndrome (SIDS). A pediatrician testified that the odds of any one such SIDS death was one in 8500, and that thus the odds of *two* such deaths, back to back as it were, equaled 1 in (8500 x 8500), that is, 1 in 73 million. There was no other substantial evidence implicating Clark.

3. Lucia de Berk worked as a nurse in the Netherlands in the 1990s. Based on a tip, the authorities found a series of six 'suspicious' deaths in a series of hospitals where de Berk had worked. She had in fact been physically present when many of the deaths occurred. A statistical expert testified the odds were about one in 342 million that these could have been coincidence. The press was merciless. de Berk had previously been a prostitute, and one of her diary entries around the time of one of the deaths noted that she had "given in to her compulsion." She was convicted of murder.

4. In 1968 in Los Angeles, witnesses described a robbery involving a black male with a beard and mustache, accompanied by a white woman with a blond ponytail. They escaped in a yellow car. The defendant, Malcolm Collins, was black, with a beard. His white female associate was blond with a ponytail. They had a yellow car. An 'instructor in mathematics' testified at trial, and offered the following probabilities for Los Angeles: black man with beard : 1 in 10; man with mustache: 1 in 4; white woman with pony tail: 1 in 10; white woman with blonde hair: 1 in 3; yellow car: 1 in 10; interracial couple in car: 1 in 1000. The prosecutor multiplied the odds and told the jury the odds were thus one in 12 million that Collins was not the perpetrator. Collins and his girlfriend were convicted.

[π] Judge Of The Superior Court of California, County of San Francisco.

5. In 1996, Dennis John Adams was charged with rape. He wasn't identified in the line up. He had an alibi, albeit from his girlfriend. But his DNA was a "one in 200 million" match to the semen taken from the victim. The jury apparently assumed this meant the odds of Adams not being the perpetrator was about 1 in 200 million. He was convicted.

6. In a similar case, Andrew Dean of Manchester, England, was convicted of rape, based primarily on expert DNA testimony that that the likelihood of the source of semen being anyone other than Dean was one in 3 million.

The logic in each of these cases seems plausible. The odds that the cases were rightly decided appear very high. But every result was faulty, premised on one, and sometimes many, confusions regarding statistics.

Below, we first explore a few traps for the statistically unwary, the better to hone our intuitions. Then we plunge to basic theory. So armed, we will return to the sorry cases outlined above. In the end, you will come to love the expression $P(H \mid E) = P(H) * P(E \mid H) / P(E)$.

II. *Warm Up Exercises: Uncommon Sense*

Perhaps, as George Bernard Shaw reputedly said, "Common sense is instinct. Enough of it is genius." Or perhaps Einstein, quoted above, was right in his dismissal. The point of course is that one never knows, in advance, whether the conclusions of 'common sense' or instinct will lead one to the right result. In the field of statistics, 'common sense' can be useful when trained to the subject, but otherwise confounds everything.

Let's begin with some warm up exercises.

1. Which of these two sequences is a random set of numbers and which devised by a human?

74329081
21234563

I invented the first sequence to make it look random; the second group is found in the infinite sequence of pi (π), at position 2,458,885.[1] Indeed, any sequence of numbers, such as "11111" (at position 32,788), "00000" (at position 17,534) and "989898" (at position 6,578), etc., can be located in π.

2. Here's a wonderful way to make a lot of money on the stock market. First, contact 100 people at random and send them a letter predicting that that the market will go up. Send a letter to another 100 people predicting the market will drop. Contact those for whom the prediction was right, and inform 50 of them that the market will go up, and 50 that the market will go down. Repeat the exercise, always simply ignoring those for whom your past prediction was false. Repeat. Finally contact those who have received accurate predictions four times in a row, and suggest they invest with you, based on your stunning track record. How can they turn you down?

[1] One counts from the first digit after the decimal point. The 3. is not counted.
http://www.angio.net/pi/piquery

Imagine you hear a company's investment vehicle has made money 5 years—make it 10 years—in a row. Would you invest? Sounds tempting. Suppose you now learn that they have 25 investment vehicles, or 1000 such investment vehicles, but are now only advertising one of them to you, the one with the remarkable track record. Still interested in investing? Not so much.

One need not invent to make the point. An advertisement in the Wall Street Journal for Dreyfus Funds noted that one of its investment vehicles, the Intermediate Term Income Fund, achieved a four-star Morningstar rating because of its remarkable performance. The ad did not mention that Dreyfus had at least 19 mutual funds. We may assume the other Dreyfus funds did not perform as well.

3. How many people do you need in a room to have a 50-50 chance that two of them will have the same birthday? Perhaps ½ * 365 or about 180 people? Not so. Only 23 people are needed.[2] This is not the same, we must recall, as a person having my specific birthday of June 18: we need 253 people to have a 50-50 chance of matching a previously specified birthday. There is a crucial point here which will be emphasized again later, which is this: while the odds *of previously specified* chance event might be very, very low—such as matching my specific birthday of June 18—the odds of *some* coincidence may be very high, such as the odds (50-50) that out of 23 people at least two of them will share a birthday.

Here's a related example. Let me specify a series of heads and tails on coin tosses:

TTHTHHHTHHHHTHTHTTH

What are the odds that you will now be able to replicate that sequence? Stunning low. Each toss is a 50-50 proposition. If we had only a *three* sequence event (e.g. T H T), the odds of a given sequence would be ½ * ½ * ½ or 1/8, i.e., 0.125. The odds of the 19 sequences event specified occurring are only 0.00000095367431640625. But I tell you that this sequence is, in fact, exactly what I obtained when I tossed my coin a few minutes ago. How is this reasonable? How is this almost incredibly rare event possible?

Similarly, I must tell you that the odds of my obtaining any given hand in bridge (13 random cards, a quarter of the 52 card pack) are less than one in 600 billion.[3] This is actually true.

Now I tell you that I was in fact dealt these cards:

♠K 10 4
♥Q J 9 6 3
♦Q 6
♣Q 8 4[4]

[2] For a visceral sense of how this is so—here we developing an *accurate* common sense—note that we are looking for the numbers of pairs of people in the room. Each pair is a possible match. With two people, there's one pair. With three, there are three pair; with 4, there are 6 pairs; with five, there are 10 pairs. And so on. By the time one gets to 23, the odds are over 50% that there will be at least one pair that matches. *See generally*, J.A. Paulos, INNUMERACY: MATHEMATICAL ILLITERACY AND ITS CONSEQUENCES (1988).
[3] Paulos, above n.2 at 40.
[4] http://playbridge.com/ Your mileage may vary. Indeed, the odds are about 1 in 600 billion that your mileage will be the same as mine.

This is also true. Incredible, is it not? It is not. While the odds of obtaining a previously specified hand such as that above are remote in the extreme, the posterior, or after-the-fact, odds of obtaining the hand I did are exactly 1—that is, it is, of course, *certain*. So too with the coin toss. After the fact, the odds that I obtained that *particular* sequence are certain; before the fact, the odds were very, very low.

So now, imagine you have just won the lottery. You are very happy, and desire to indulge your fantasies: you are carrying a million dollars in cash. "I can make paper hats from this money," you mutter, "I can wear money hats." Your reverie is interrupt by the police: they arrest you for bribery; or theft (take your pick). You protest that the money is from the lottery. They sneer, and the cuffs come out. The police know very well the odds of winning the lottery are (for example) one in 175,711,536,[5] so they conclude this innocent explanation is *most* unlikely. Indeed, these odds are far, *far* worse than those in the Clark, Collins and Dean cases above, and are about what they were in the Adams case. And you know what happened to them.

Our statistical intuition, now increasingly attuned, smells a rat. We will come back to this.

Another related phenomenon is known as the "sharpshooter effect," and it too preys on the confusion between posterior (after the fact) and anterior (before the fact) probabilities. Here's the set up. A man responds to an ad in the magazine *Soldier of Fate,* applying for the assassin's job you have advertised. He hands you a true photo of a rather small target painted on a large piece of wood, all the bullet holes beautifully clustered about the center of the target. He says, truthfully, that he actually hit as indicated from, say, 1000 yards using his M24 Sniper Weapon System with a Leupold Mk.4 LR/T M3 10x40mm first focal plane fixed-power scope with a mil-dot reticle. Do you accept his expertise, and hire him? Wait. This is what you do not know: He *first* used his M 24 to shoot a hundred rounds at a barn door, and *thereafter* painted the target over the best cluster he could find.

I take it you will not offer him the job.

The sharpshooter effect can be more insidious. Diseases such as cancer will cluster, just as numbers ("0000" or anything else) in π will cluster. Given a large enough area—the equivalent of the barn door—one will find cancers spread out in some areas and concentrated in others, merely as a result of random chance. *There must be a reason for the cluster*, one thinks, but that thinking is no better than assuming my would-be assassin can hit a real target. Inevitably the cluster will correlate with *something*: a nearby factory, telephone wires, very high (or very low) income, a high concentration of fast foods restaurants, a nearby army base, etc. But the apparent after-the-fact correlation is meaningless,[6] as empty as one's surprise at learning that two people in

[5] http://en.wikipedia.org/wiki/Lottery#Probability_of_winning

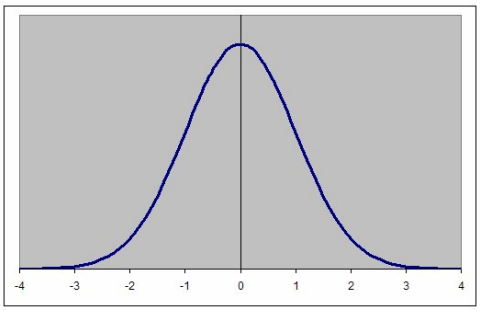
[6] "A community that is afflicted with an unusual number of cancers quite naturally looks for a cause in the environment – in the ground, the water, the air. And the correlations are sometimes found: the cluster may arise after, say, contamination of the water supply by a possible carcinogen. The problem is that when scientists have tried to confirm such causes, they haven't been able to. Raymond Richard Neutra, California's chief environmental health investigator and an expert on cancer clusters, points out that among hundreds of exhaustive, published investigations of residential clusters in the United States, not one has convincingly identified an underlying environmental cause. Abroad, in only a handful of cases has a neighborhood cancer cluster been shown to arise from an environmental cause. And only one of these cases ended with the discovery of an unrecognized carcinogen." *The Cancer Cluster Myth*, The New Yorker, Feb. 1999.

a class of 23 share a birthday. With an infinite supply of things to correlate with, *some* correlation will always be found.[7]

There are other terms denoting this kind of bad thinking. Cherry picking. Data dredging. As one researcher has shown, one might collect a few years' data about trout survival, decide to see what it might correlate with—dredging for data—and eventually find a statistically 'valid' result that it correlates with cold winters.[8] Cold winters, more trout. That result is gibberish unless and until the hypotheses is tested on different set of data not used in the initial dredge, for example, on trout from other lakes and other years.

4. Let us look at correlation and causation. This is more obvious: We know that the correlation of (i) the rooster crow at dawn and (ii) sunrise does not imply that the rooster caused the sun to rise. We know that two correlated events (Y, X) may be related in a variety of ways: X may cause Y; Y may cause X; and Z might cause both X and Y without Y and X having any effect on each other.

Or not. There may be no link *at all* between the correlated events. A wonderful story is told of an air force instructor's reaction to the behavior of his trainee pilots. He berates and chastises them after lousy landings—and lo! Their landings improve thereafter. He complements them after first rate landings- and lo! Thereafter the landings get worse. The instructor concludes that he must never complement, and only chastise. Is he right? He is not. The instructor has merely come across the notion of a *regression to the mean*. Think of all the landings one might make- good, bad, indifferent. Most will be alright, in the middle. Some—not too many--will be very, very good; and some—not too many—will be terrible. The old bell curve tolls again, with most of the events clustering around the middle:



very good *ok* *very bad*

After a stunningly good landing, which type of landing is most likely next? One *not* as good, something towards the center of the graph. So too with a remarkably bad landing—the next one is most likely to be better, with or without chastisement, complements, or any other communication from the instructor. The landings regress to the mean. The same thing is found with heights of children: critical genetic abnormalities aside, very tall parents will have somewhat shorter children, and the children of very short parents will tend to be taller than their parents. The heights regress to the mean.

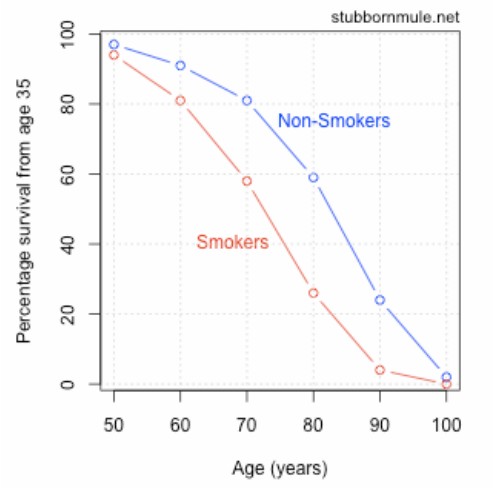
Regression to the mean will support false conclusions from data which appears statistically valid. For example, if one measures very high blood pressure, and then treats, one might then find lower

[7] *See generally*, W. Thompson, "Painting The Target Around The Matching Profile: The Texas Sharpshooter Fallacy In Forensic DNA Interpretation," 8 *Law, Probability and Risk* 257–276 (2009).
[8] http://warnercnr.colostate.edu/class_info/fw663/dredging.pdf.

pressure. This suggests the efficacy of the treatment. But if subjects are examined later without treatment, one may find that high blood pressure *still* declines, due to regression towards the mean. The treatment had, in fact, no impact.[9]

In the same vein I note a study which looked at recidivism. An official argued that prison sentences were effective because, after release from prison, the offence for which ex-convicts were convicted were usually for a less serious crime than the original one. But as a group, denizens of prisons tend to have committed crimes at the more serious end of the criminal spectrum, and so, generally speaking, if they commit a new crime it will be of less severity, simply as a function of regression to the mean.[10] Prison may or may not be effective under various circumstances, but this data doesn't tell us.

5. Now we move to the wonderful world of graphs, charts, plots projections and induction. The underlying problems is very similar to those explored above. Data is charted, a line drawn to "best fit" the data, and extrapolations or projections are made. The underlying assumption is that the data reflects a real correlation, so that for example, we might find something like this, plotting cigarette consumption and death dates, suggesting a correlation between the two.



But the fact that a graph can be drawn does not suggest its explanation has any meaning. The drawing of a graph is no more than the expression of an apparent correlation, and so is no more meaningful than the underlying claim of correlation. I can plot cancer against incidents of swimming pools in Las Vegas, or show you a "direct statistical relationship between pig iron production in the United States and the British birth rate,"[11] but it won't tell you anything.

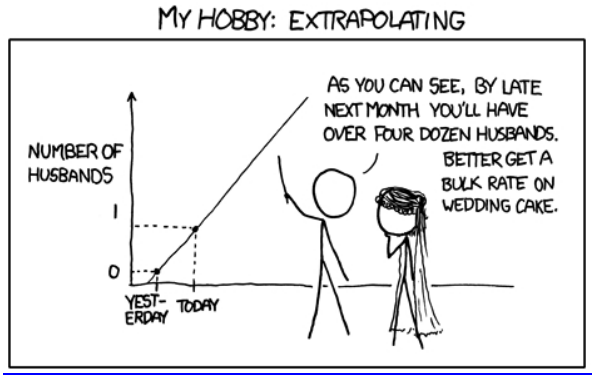
And because charts imply patterns, they imply extrapolation. But this may be wholly fallacious. The plot of the stock market up to October 1929 was not predicative of the subsequent market, nor would it have been predictive to plot the market up to the dot com bubble or housing prices up to 2007.

---

[9] http://www-users.york.ac.uk/~mb55/talks/regmean.htm
[10] http://www-users.york.ac.uk/~mb55/talks/regmean.htm
[11] M, Thompson, *Comment*, "Causal Inference In Epidemiology: Implications For Toxic Tort Litigation," N.Car.L.Rev. 249 (Nov. 1992).

Finally, even accurate chats—that is, charts which actually reflect causal nexuses—can be complex and suggest false conclusions when the wrong part of the chart is viewed, or the chart is viewed at too high or too low a detail level. The plot of alcohol consumption against various diseases shows different things at different points. For example, it shows an initial decrease in risk, perhaps suggesting a beneficial threshold effect, but then shows a large increase in risk with increase intake of alcohol.

A favorite chart:[12]



The study of misleading charts is an expertise all its own, and the subject of first class treatments.[13]

6. A last exercise to develop our statistical common sense. The scene is your doctor's office. Assume a fatal disease affects about 1 out of every 1000 people.[14] There's a test: it has a zero false negative rate and a 5% false positive rate. That means that if a person actually suffers from the disease, the test always gets it right; and if a person does not have the disease, 5% of the tests will nevertheless indicate they do have the disease. The doctor gives you this news: You tested positive for the disease. How do you feel? Most people (including doctors by the way[15]), conclude that the odds that you have the disease are about 95%. Five percent false positive or "failure" rate for the test, after all, right? Time to check the will and make funeral arrangements?

Probably not. The odds of having the disease we know are 1 in 1000; if the test is given to 1000 people, one person will test positive, accurately. Five percent of the tests—tests administered to 50 people—will also show positive for the disease—but they don't have it. You are one of the 51

[12] Used with permission, http://xkcd.com/about/.

[13] The master is Edward Tufte. *See e.g.*, E. Tufte, THE VISUAL DISPLAY OF QUANTITATIVE INFORMATION (1983); BEAUTIFUL EVIDENCE (2006); *see generally* http://www.edwardtufte.com/tufte/. Tufte concentrates more on misleading, incomprehensible graphical presentation of facts, rather than the sort of statistical fallacies I have alluded to, but his insights apply. See especially his chapter in BEAUTIFUL EVIDENCE at 141, called "Corruption On Evidence Presentations: Effects Without Causes, Cherry-Picking, Overreaching, Chartjunk, and the Rage to Conclude." There is much literature on the subject of the dangers of bad graphical presentations; this material goes way back. I have a 24th edition of the 1954 classic, D. Huff, HOW TO LIE WITH STATISTICS.

[14] For example, Alzheimer's disease. This affects about one in 1,000 people between the ages of 30 and 49. http://www.huffingtonpost.com/brandon-colby-md/alzheimers-waging-a-new-o_b_574002.html

[15] Doctors do medicine, not statistics. Sally Clark, convicted on a pediatrician's 'expert' testimony, would have wished the court had insisted on the distinction. For more on doctors' errors calculating the correct probabilities, *see e.g.,* Peter Sedlmeier & Gerd Gigerenzer, "Teaching Bayesian Reasoning in Less Than Two Hours," 130 *Journal of Experimental Psychology: General* 380-400 (2001).

people, but the odds are 50/51 that you don't have the disease. The odds are, that is, about 2% that you have the disease. Not 95%.

This intuitive error is known as base rate negligence, discounting the base rate, or more commonly the base rate fallacy. One's intuition has not accounted for the decisive impact of the base rate-- the fact that the disease is rare, the odds that that a random person would have the disease. Even with e.g., a 99% accurate test, when a disease is rare, the number of false positives will vastly outweigh the accurate tests.[16]

By way of a preview of where we're going with this, let us apply this example to a simple DNA match problem. The defendant is a match for the DNA at the scene of the crime. The odds of such a genetic match are one in a billion. Should the jury convict, happy in the thought that the odds that the verdict is wrong to be about one a billion? No. In a world population of seven billion, there are seven people who perfectly fit the DNA. All other evidence aside, the odds of the defendant being the culprit are actually 1 in 7.

But of course, we usually do have other evidence. Perhaps an item left at the scene, an eye witness description of a man race, or a woman's pony tail. Our intuition now properly cleansed and alert, we move into the dark heart of probability, and work with a powerful theorem to calculate odds of an unknown, given what we do know (such as a fingerprint, the ponytail description, or other evidence).

III. *Thomas Bayes and Conditional Probability*

Given something we know (e.g., woman has blond hair), what are the odds of something else (blond Marilyn was at the scene) being true? The critical thing here is to realize that this question is not the same as asking, given Marilyn was at the scene, what the odds are of her being a blond. An expression might help: Odds of <A given B> ≠ <B given A>. So, the odds that an English speaker is an American, is, say, 1/5. But the odds that an American is an English speaker is more like 0.95.[17] Given a judge, what are the odds the person went to law school? In California, 100%. Given a law school education, what are the odds the person is a judge? About .01, one out of every 100.[18]

The Reverend Thomas Bayes was an Eighteenth century English minister and mathematician who concerned himself with what folks at the time termed inverse probability problems. Probability problems were of this nature: from an urn of X black and Y white balls, what are the odds I will

---

[16] Suppose a test to spot terrorists is 90% accurate—this means almost nothing. http://news.bbc.co.uk/2/hi/uk_news/magazine/8153539.stm. See also N. SINGER, "In Push for Cancer Screening, Limited Benefits," The New York Times (July 16, 2009), http://www.nytimes.com/2009/07/17/health/17screening.html?_r=1 ("A recent European study on prostate cancer screening indicated that saving one man's life from the disease would require screening about 1,400 men. But among those 1,400, 48 others would undergo treatments like surgery or radiation procedures that would not improve their health because the cancer was not life-threatening to begin with or because it was too far along. And those treatments could lead to complications including impotence, urinary incontinence and bowel problems.")

[17] Paulos at 63-64. The odds might be somewhat less these days, say 82% or 0.82. http://wiki.answers.com/Q/What_percentage_of_American_citizens_speak_english_as_a_first_language.

[18] http://members.calbar.ca.gov/search/demographics.aspx (227,952 lawyers); http://www.courtinfo.ca.gov/qna/qa7.htm (about 2,280 judges on all state courts).

draw a white ball? Inverse probability asked, (to use the phraseology used just above), given that I draw a white ball, what is the likely proportion of white and black balls in the urn?

Here, I'll introduce Bayes' famous theorem. It formalizes some of the thinking discussed above (and is a resource for long winter nights in upper Norway). I'll then move onto another more natural way of thinking about Bayesian probabilities, already forecast by the work above.

Here's the classic formula:

$$P(H \mid E) = P(H) * P(E \mid H) / P(E)$$

E is evidence. H is the hypothesis you're interested in. P is probability (or the odds), but as you'll see it is used in two different ways in the classic formula: P means, on the *left* side of the equation, the *posterior* probability, that is, the *final* odds that the hypothesis you're interested in is true. On the right hand side of the equal sign, P is the *prior* probability that the hypothesis is true, i.e., your *initial* working assumption about the truth of the hypothesis, as it were. The sign "|" means "given that," as I've used the phrase above. So H|E means "H given that E is true." The stuff on the right hand side of | is what you know, the stuff to the left of | is the hypothesis you are concerned with. I might be concerned about the odds that, given I have two children, I might go mad. I would phrase this so: "go mad | has 2 children."

So the formula reads as follows, in something closer to the Queen's English: the *posterior* probability of the hypothesis, given the new evidence [P(H|E)] (this is what we're interested in) equals the following: (1) the *prior* odds or probability of the hypothesis [P(H)] multiplied by (2) the probability of the evidence given the hypothesis [P(E | H)], all divided by [/] the probability of the evidence [P(E)]. The *prior* probability is the state of our knowledge or ignorance concerning the hypothesis *before* we have taken into account the new evidence E, and the formula thus in effect tells us the impact of the new evidence [E] on our initial thoughts about the odds of H being true.[19]

Let's apply the formula. Assume these things[20]:

- The probability that a woman who undergoes a mammography will have breast cancer is 1%.
- If a woman undergoing a mammography has breast cancer, the probability that she will test positive is 80%.
- If a woman undergoing a mammography does not have cancer, the probability that she will test positive is 10%.

Now, what is the probability that a woman who has undergone a mammography actually has breast cancer if she tests positive? Once again, Bayes' formula:

$$P(H \mid E) = P(H) * P(E \mid H) / P(E)$$

The hypothesis that we're testing here is that the woman has cancer. The new evidence we're evaluating (E) is the positive test. P(H | E) expresses what we're interested in, the odds (or posterior probability) of the hypothesis (i.e., she has cancer) *given* a positive test.

---

[19] A nice explanation is found at R. Posner, HOW JUDGES THINK 66 (2008).

[20] Peter Sedlmeier & Gerd Gigerenzer, "Teaching Bayesian Reasoning in Less Than Two Hours," 130 *Journal of Experimental Psychology: General* 380-400 (2001).

P(H) is the odds of the hypothesis being true. We know from the facts stated the answer to this: it's 1%, or 0.01.

P(E | H) is the odds that the evidence is true, given the hypothesis. We know this too from the stated facts: if a woman has cancer (the given here) then we'll have the evidence (the positive test) in 80% of the cases, that is, 0.8.

P(E) is the probability of the evidence, that is, the odds that the test will be positive. We can figure this out, as well. We know that the 1% of woman who have breast cancer test positive 80% of the time, and the 99% of women who don't have cancer test positive 10% of the time. So the calculation (.01 * .8) + (.99 * .1) should total all the positive tests, i.e., .008 + .099 = 0.107.

Now we can plug these figures in and we get:

$$P(H \mid E) = 0.01 * 0.8 \ / \ 0.107 \text{ or } 0.0747.$$

So the result is, given the positive test, the woman has a .075 chance of *actually* having cancer. That's the *posterior* (i.e., after the calculation, taking into account the new evidence (E) of the positive test) probability.

Aside from the raw power of Bayes' theorem, it uses—and very clearly distinguishes—two expressions which are the subject of deep confusion in most of the six cases I described in the opening segment of this Note. This is, I think, the dark heart of the horrific errors in those cases that led to unjustifiable convictions and prison terms. These are the two terms:

$$P(H|E) \text{ and } P(E|H)$$

Recall what they mean. The first is the probability of the hypothesis given the [new] evidence. The second is the probability of the evidence given the hypothesis. If you mix these up, the analysis is complete rubbish. In the cancer example I have above, the two expression were, respectively, (i) the probability that the woman has cancer, given the positive test, as contrasted with (ii) the probability of a positive test, given she actually has cancer. We know now that these two probabilities vastly differ in their meaning, and in their number. P(H|E) was 0.0747, and P (E|H) was 0.8.[21] In a DNA context, where forensic experts are tossing around numbers, P(H|E) is the probability that someone's DNA matches the DNA at the scene; P(E|H) is the probability that someone who is a match is actually innocent. The expert usually provides the first figure; but the jury believes it has been given the second one. This confusion is the *prosecutor's fallacy*:

> The prosecutor's fallacy is the assumption that the random match probability is the same as the probability that the defendant was not the source of the DNA sample. See Nat. Research Council, Comm. on DNA Forensic Science, The Evaluation of Forensic DNA Evidence 133 (1996) ("Let P equal the probability of a match, given the evidence genotype. The fallacy is to say that P is also the probability that the DNA at the crime scene came from someone other than the defendant"). In other words, if a juror is told the probability a member of the general population would share the same DNA is 1 in 10,000 (random match probability), and he takes that to mean there is only a 1 in 10,000 chance

[21] Similarly, when I first introduced Bayes' formula above I gave an example of P(H|E) as "go mad |has 2 children" i.e. given that I have 2 children, what are the odds I will go mad? Perhaps the odds here are very high indeed. This is the very different from asking, given that I am mad, what are the odds that I have two children? If madness is caused by a thousand things, the odds of madmen having children may be low.

> that someone other than the defendant is the source of the DNA found at the crime scene (source probability), then he has succumbed to the prosecutor's fallacy. It is further error to equate source probability with probability of guilt, unless there is no explanation other than guilt for a person to be the source of crime-scene DNA. This faulty reasoning may result in an erroneous statement that, based on a random match probability of 1 in 10,000, there is a .01% chance the defendant is innocent or a 99.99% chance the defendant is guilty.
>
> *McDaniel v. Brown*, 130 S.Ct. 665, 670 (2010).

There's a more intuitive way to figure out Bayesian probabilities. It's termed the *natural frequency* approach,[22] but I think of it as the *counting-with-fingers* approach. It works for me, and indeed I used the counting technique to check my work above. You have already seen this counting approach: I also used it above to explain the base rate fallacy. Here it is again, using the mammogram numbers I used to discuss the formal Bayes theorem.

Recall the stated facts:

- The probability that a woman who undergoes a mammography will have breast cancer is 1%.
- If a woman undergoing a mammography has breast cancer, the probability that she will test positive is 80%.
- If a woman undergoing a mammography does not have cancer, the probability that she will test positive is 10%.

So now we simply—and carefully—count out what all this means. Let's start with 1000 women. Given the stated facts, we know this: 10 have cancer. Of that number, 8 will show up with positive tests. We also know that 990 (1000 – 10) women do not have cancer. But we also know that 10% of them will *still,* nevertheless, have positive tests. So that's 99 positive tests for the actually healthy women.

We now know how many total positive tests there are: 8 + 99 = 107 positive tests. We can now calculate the odds that a positive test really shows cancer: Eight of 107 positive testers really have cancer, i.e.. 8/107, which is .0747. That's the same number we got with the formal theorem.

IV. *Reprise*

> "Now with gimlet eye do we re-entrain our stories."

1. We turn back to professor Dershowitz, who in a moment of partisan enthusiasm—well, not so momentary[23]—suggested the odds were low that O.J. Simpson had killed his wife, because fewer than 1 in 1000 (he also later suggested the odds were 1 in 2,500) abused wives are later killed by their abusers. But we are not, never were, concerned about the odds of the wife being killed given that she was abused. Those odds may well be very low. We already *know* she was killed.

---

[22] Peter Sedlmeier & Gerd Gigerenzer, "Teaching Bayesian Reasoning in Less Than Two Hours," 130 *Journal of Experimental Psychology: General* 380-400 (2001); S. Strogatz, "Chances Are," Opinionator column, *The New York Times* (April 25, 2010).

[23] Dershowitz repeated variations of his error in letters to newspapers, TV interviews, and in a book he wrote in 1997, some years after the trial. See generally http://isds.bus.lsu.edu/chun/teach/reading-a/ojsimpson.htm

So we ask: If a man abused his wife and she is later murdered, what are the odds the abusing spouse was the murderer? About 80-90%.

Assume 100,000 battered women: if Dershowitz is right, 40 of these will be murdered by their batterer. If the annual murder rate or all women in the county at the time was 1 in 20,000[24]—and here we're invoking the base rate, which Dershowitz does not—then we add in another 5 of these women in as being killed by someone else. So out of a total of 45 murdered women, 40 of them were killed by the batterer. That's about 90% of the time.

2. And on to Sally Clark. Recall the expert pediatrician (not statistician) argued that the odds of a single SIDs death was 1 in 8500; and thus if two such deaths, the odds would be 8500 * 8500, or 1 in 72,250,000. It *is* correct to estimate the odds of two independent events by the product of their individual odds, so that the odds of *two* unrelated dice turning up "6" is 6 * 6 or 1 in 36. That is what this "expert" did. But the first error here was in assuming independent events—the truth is, as a matter of fact, otherwise: there are likely genetic or other common reasons for SIDS death, so that if 1 child dies of SIDS, the odds of another dying from SIDS may be about 1 in 60 or 1 in 100,[25] suggesting the odds of two such deaths in a family to be one in 130,000. Given the population of England, we are almost *guaranteed* to see double SIDS deaths from time to time without the commission of a crime, just as we are guaranteed to see cancer clusters from time to time.

SIDS death rate in England and Wales in 1993-98 were about one in 1,785 (the number used at Sally Clark's trial, the much higher one of 8600, was putatively to account for Clark's age and social status). Because the odds of two SIDS deaths are not independent, but appear to be that the second child has roughly twice the normal chance of dying from SIDS if the first child did, we may calculate the odds of the first death at 1 in 1875; the second as 2/1875, and the combination at 1 in 1,593,112. With a population of about 3 million families in England and Wales with two or more children, we will see around two cases of double SIDS deaths per year.

Clark was convicted on this inadmissible evidence (one in 73 million) in 1999; her first appeal was denied although the judges noted the problem; a second appeal was granted (as they say in the U.K.) and she was freed from prison in early 2003. She died four years later at the age of 42. Her family said she never recovered from her ordeal. The BBC reported "Post-mortem tests showed she had a concentration of alcohol in her blood which would have made her five times over the drink drive limit."[26]

3. Nurse de Berk, we were told was guilty because there was a 1 in 342 million chance that the six deaths at which she was present could have been mere coincidence. The mere phraseology of this conclusion now should make you wince. Let us trot out again the two terms, which were mixed up in the prosecution's case:

$$P(H|E) \text{ and } P(E|H)$$

We have been told an estimate of the odds that so many deaths could have occurred with de Berk present if she was innocent—the odds of the innocence hypothesis (H) given the death evidence (E)—in short, (P(E|H)). *But that's not the right question*. Actually, we want to know the odds

[24] This is from the Opininator, above at n.22 as is the balance of the analysis here.
[25] http://www.sallyclark.org.uk/observer0107.html
[26] http://news.bbc.co.uk/2/hi/uk_news/england/essex/7082411.stm

that de Berk is innocent (H), given the evidence of the deaths (E)—in short, P(H|E). The prosecutor's fallacy again.

De Berk was given a life sentence for multiple murders in 2004. As the nature of error came to light, she was released in 2008 pending appeal, and was exonerated just recently, in April of 2010.[27]

4. If you enjoyed the multiplication of 8600 with 8600 in the Clark trial, you will find particularly delicious the multiplication in the Collins case. Recall the 'expert' there multiplied the odds of things such as 'black man with beard' and 'man with mustache' and white woman with pony trial' and 'white woman with blond hair' as well as 'interracial couple,' and so on. These factors were treated as if each were independent of the others, as are the rolls of multiple dice. But of course that is absurd: the 'interracial couple' is not only not independent of the 'white woman' and 'black man', it's the very same thing. And 'white woman' is not independent of 'blond hair,' but closely related. And so on. Yet the Collinses were convicted and sentenced to prison.

The state Supreme Court reversed. *People v. Collins*, 68 Cal. 2d 319 (1968). First, it noted that there was actually no evidence to support the factors—that is, the prosecution never showed that the odds were in fact, that only one out of every ten cars which might have been at the scene of the robbery was yellow, or that one in four men have a mustache, or one in every ten women has a ponytail. They just made the numbers up. The Court then went on to note that with Los Angeles' large population, even if one accepted the prosecution's invented numbers, the odds were about 40% that another couple matching the description existed, which in turn makes the odds of the Collinses being guilty closer to 50-50 than 1 in 12 million (or, as the prosecutor actually argued in closing, "something like one in a billion"). Glumly, the Court noted that "few defense attorneys, and certainly few jurors, could be expected to comprehend this basic flaw in the prosecution's analysis. Conceivably even the prosecutor erroneously believed" his theory, the court remarked.

5. John Adams was convicted of rape because the jury inferred that, because the match rate of the semen was 1 in 200 million, the odds of Adams being guilty were the same. (As I note below, the defense argued this number was wrong.) But the victim said Adams did not resemble her attacker; his age was substantially different from that she reported, and he had an alibi which was not broken. The prosecution had no evidence, at all, except the DNA.

You know the basic fact required to put this number (1 in 200 million) in the proper context. It is the number of men in the relevant population. In a population of 100,000 men, the odds would be 100,000 in 200 million, or 1 in 2,000 chance that someone else is a match as well. In this case, the defense team knew all about Bayes, and actually tried to educate the jury, all to no avail. Again, the forensic evidence was in effect the probability of finding that DNA sequence if the semen had come from someone else (say, one in 200 million), whereas the evidence was understood as the probability that the semen was left by Adams. Now, even correctly understood, the Bayesian probabilities on their surface painted a grim picture. Assume the male population of England is 25 million. The number of men therefore that also could have been the attacker = 25/200 or .125 men. This 'fractional' man might not seem to any jury to raise much of a reasonable doubt. Using the defense number for match probability of one in two (not 200)

[27] http://www.dutchnews.nl/news/archives/2010/04/nurse_lucia_de_berk_not_guilty.php; http://www.independent.co.uk/opinion/commentators/nigel-hawkes-did-statistics-damn-lucia-de-berk-1940735.html

million, the result is far better: given 25 million men, there are probably 12 other men with the same DNA signature.

But even if we accept the prosecution's case, the result is suspect, because there is one more fact: The 1 in 200 million match number ignored the possibility that the attacker was a close relative of Adams: a critical issue, since the defendant had a brother whose DNA was never tested.

6. We can now provide a proper context to the Dean case as well. If the match probability were truly one in 3 million, then with a population of 25 million men, it is probable that about 8 men match the profile; so one might say the odds of Dean being guilty are only 1/8. Telling the jury, as the prosecutor did in the trial, that the odds that the perpetrator being anyone other than Dean were one in three million is entirely wrong. Dean's conviction was reversed on appeal.

In fairness, we must also remember that the base rate may not really be a function of all 25 million men in the United Kingdom, because not every one of them had the opportunity to commit the crime. The facts of a case might suggest a small population. The population of a walled in, locked city might be only 100,000; or more reasonably, perhaps only a certain number of men were within travel distance of the scene of the crime. However the matter is handled, the important point is that the jury must understand the assumptions that underlie the calculations.

V. *An Aside on DNA*

The statistical issues discussed above most often arise in the area of DNA results, so I have thought it worthwhile to approach the end of this Note with cautionary tales specific to this area.

In the literature, the *Collins* case is generally cited for the basic lesson I provided: The risks of mixing up dependent and independent factors, and the base rate fallacy. But I highlighted another issue, as did the California Supreme Court: the numbers used by the prosecution as *predicates* for their idiot calculations were themselves suspect (actually, in *Collins*, just made up).

This problem lurks in DNA analyses. That is, while the analyses are now so sophisticated and count so many distinctions among human DNA that match figures such as "one in a billion" are possible, the underlying analyses may be subject to varying interpretations. Different laboratories will count different specific genetic sequences (alleles). Some will count perhaps 11 of them, some 13. When the DNA of multiple persons are located, judgments must be made whether alleles belongs to one person or another. One researcher provided the same sample to 69 laboratories and obtained results which differed by 10 orders of magnitude.[28] One order of magnitude is the difference between 1 and 10. Ten orders is the difference between 1 and 10,000,000,000. The difference in results is catastrophically enormous: Assume a population of 1 million eligible people, and the first lab gave a match result of one in 10,000: we'd have 100 people whose DNA matched the specimen. The second lab, with results 10 orders of magnitude distant, would give a match probability of $1/10^8$, an infinitesimally small number strongly suggesting only the suspect was a match.

In another exercise, the DNA of a suspect was compared to that at the crime scene and the original laboratory concluded the defendant 'could not be excluded' as the perpetrator.

[28] L. Geddes, "How DNA Evidence Creates Victims of Chance," New Scientist (18 August 2010). See also L. Geddes, "Fallible DNA Evidence Can Mean Prison or Freedom," New Scientist (11 August 2010).

Seventeen other labs were then given the same samples. One agreed with the original analysis. Four reported inclusive results. Twelve labs said he *could* be excluded.[29]

The results above are from laboratories honestly trying to do the right thing, given samples by people with no ulterior motives. It may not always be possible to gamble on that. Because DNA can in effect be scrubbed and external DNA can be inserted, it's now possible to fake DNA evidence to make it match a specific person.[30]

VI. *Conclusion*

"A one in a million" shot, someone says admiringly of a particularly marvelous golf swing, say a hole-in-one. Not so marvelous. In the United States with 300 million people, the event will occur roughly 300 times. "Incredible" coincidences happen all the time, but mean nothing.

Let us not exaggerate, though, because while it is true that any apparent 'pattern' can be located in an infinite sea of actually random data, our contexts do not generally involve an infinity of people or time. Thus, to suggest that a group of typing monkeys will eventually write Hamlet is true only in a universe that lasts forever, which is, I am afraid, probably (I use the word advisedly) not our universe.[31] And just because we can find any string of numbers in the endless π does not mean we are assured of finding any particular event or sequence in our own lives. As we saw with the base rate fallacy, everything depends on the size of the universe: without that number, probability estimates are gibberish; with it, those estimates may show a truly high or low probability of an event.

There are a raft of other statistical fallacies I have not explored. We have the 'ecological fallacy', whereby conclusions about groups are applied willy-nilly to individual members of the group. Just because, for example, the odds of SIDS in households with no smokers, at least one wage earner, and a mother over 27 are generally, say, 1 in 8600, it does not mean those were the odds in Sally Clark's household. A brick is not an example of a house. The 'gambler's fallacy' is a variant of those we've examined above—ten heads in a row suggests high odds of tails next time, our imaginary gambler believes. But not so, of course. The "law of averages" does not say anything about a specific event (such as the coin toss), any more than it predicts what the 1,56,987th number in π will be. We have 'selection bias' or sample bias in which the universe of items or events or people subject to comparison already predisposes the result. So, for example, we might test a drug against cancer on people under 30, and see a terrific result because the cancers suffered by the young are very different from those that afflict the elderly—and most cancers affect the elderly.[32]

Many reports, especially those that reach us through the popular media—or advertising—leave out critical information needed to evaluate its relevance. Let me give you an example. Studies, such as polls, give us results with a certain confidence level and a margin or error. So we might see a poll telling us, with a 95% confidence level, that Genghis Khan is leading Mao Tse Tung by 3 points, with a plus or minus (±) 3 margin of error. The same data may support a higher

---

[29] L. Geddes, "Fallible DNA Evidence Can Mean Prison or Freedom," New Scientist (11 August 2010).
[30] John Trimmer, "CSI Fraud: Researchers Craft Fake DNA Evidence" (August 18, 2009) http://arstechnica.com/science/news/2009/08/dna-samples-used-by-crime-labs-faked-in-research-lab.ars
[31] http://en.wikipedia.org/wiki/Infinite_monkey_theorem
[32] *See generally*, B. Goldacre, BAD SCIENCE: QUACKS, HACKS, AND BIG PHARMA FLACKS 158 *et seq.* (2008-2010). The example in the text is my own, not Goldacre's.

confidence level, say 99%, but with a greater margin of error, say ± 15. Or we might have a lower confidence level of 75% with a very small margin or error, ± 1. Give all the options, Mao Tse Tung might even be reported as leading Genghis Khan with a 32% confidence level and a ± 68 margin or error—complete gibberish, of course.[33] Unless you have both numbers—the confidence level and the margin or error—you have nothing. Think about the last poll you read. You had nothing.

Let me end with a sad story. The renown British Broadcasting Service, the BBC, reported results on Britain's happiest and most miserable cities. Exquisite Edinburgh was reported as the most miserable place in a study undertaken by two prestigious universities. But buried deep, deep in the story, and not reported in the many local papers that ran the item, was a little comment: "the researchers stress that the variations between different places in Britain are not statistically significant." They interviewed 5,000 people in 274 areas, which allows for less than 20 per locale. The reports were just random noise.[34]

We are not sad, then, for the unhappy life of those in Edinburgh. They have, after all, haggis and bagpipes. We are only sad at the incontinent use of pretend statistics, which can be funny when we see them in the media, but not funny, at all, in court.

π

[33] Anything less than 95% confidence level is usually pointless.

[34] The story is from Goldacre's website, which contains a host of other wonderful stories and discussion. http://www.badscience.net/2008/09/the-certainty-of-chance/#more-788. In point of fact, we do not need zillions of people in our sample to get a reasonable number. Assuming the question calls for basically only a yes/no response (e.g., are you happy?), and that the City's population to be 448,624, as it was in 2001, and we are content with a confidence level of 95% (we're "95% sure we right") and accept a confidence interval of 5 (so the numbers are good within ± 5), we need interview only 384 people. With only 20 people in the sample, the confidence level drops to around 20, which means that if the results were that 30% of the City was unhappy, the real number is somewhere between 10% and 50%, at best. The real import of that, in turn, depends on how the *other* cities fared, for if most of them also had reported results anywhere in the 10%-50% area, then no comparison can be made. Have fun yourself with the calculator at http://www.surveysystem.com/sscalc.htm.