11

BAD STATS

Now that you appreciate the value of statistics—the benefits and risks of intuition—we can look at how these numbers and calculations are repeatedly misused and misunderstood. Our first examples will come from the world of journalism, but the true horror is that journalists are not the only ones to make basic errors of reasoning.

Numbers, as we shall see, can ruin lives.

THE BIGGEST STATISTIC

Newspapers like big numbers and eye-catching headlines. They need miracle cures and hidden scares, and small percentage shifts in risk will never be enough for them to sell readers to advertisers (because that is the business model). To this end they pick the single most melodramatic and misleading way of describing any statistical increase in risk, which is called the relative risk increase.

Let's say the risk of having a heart attack in your fifties is 50 percent higher if you have high cholesterol. That sounds pretty

bad. Let's say the extra risk of having a heart attack if you have high cholesterol is only 2 percent. That sounds OK to me. But they're the same (hypothetical) figures. Let's try this. Out of a hundred men in their fifties with normal cholesterol, four will be expected to have a heart attack, whereas out of a hundred men with high cholesterol, six will be expected to have a heart attack. That's two extra heart attacks per hundred. Those are called natural frequencies.

Natural frequencies are readily understandable, because instead of using probabilities, or percentages, or anything even slightly technical or difficult, they use concrete numbers, just like the ones you use every day to check if you've lost a kid on a bus trip or got the right change in a shop. Lots of people have argued that we evolved to reason and do math with concrete numbers like these, and not with probabilities, so we find them more intuitive. Simple numbers are simple.

The other methods of describing the increase have names too. From our example above, with high cholesterol, you could have a 50 percent increase in risk (the "relative risk increase"), or a 2 percent increase in risk (the "absolute risk increase"), or, let me ram it home, the easy one, the informative one, an extra two heart attacks for every hundred men, the natural frequency.

As well as being the most comprehensible option, natural frequencies contain more information than the journalists' relative risk increase. Recently, for example, we were told that red meat causes bowel cancer, and ibuprofen increases the risk of heart attacks; but if you followed the news reports, you would be no wiser. Try this, on bowel cancer, from the *Today* program on Radio 4: "A bigger risk meaning what, Professor Bingham?" "A third higher risk." "That sounds an awful lot, a third higher risk; what are we talking about in terms of numbers here?" "A difference . . . of around about twenty people per year." "So it's still a small number?" "Umm . . . per 10,000 . . ."

These things are hard to communicate if you step outside the simplest format. Professor Sheila Bingham was the director of the MRC Centre for Nutritional Epidemiology in Cancer Prevention and Survival at the University of Cambridge and dealt with these numbers for a living, but in this (entirely forgivable) fumbling on a live radio show she was not alone; there are studies of doctors, and commissioning committees for local health authorities, and members of the legal profession that show that people who interpret and manage risk for a living often have huge difficulties expressing on the spot what they mean. They are also much more likely to make the right decisions when information about risk is presented as natural frequencies, rather than as probabilities or percentages.

For painkillers and heart attacks, another front-page story, the desperate urge to choose the biggest possible number led to the figures being completely inaccurate in many newspapers. The reports were based on a study that had observed participants over four years, and the results suggested, using natural frequencies, that you would expect one extra heart attack for every 1,005 people taking ibuprofen. Or as the *Daily Mail*, in an article titled "How Pills for Your Headache Could Kill," reported: "British research revealed that patients taking ibuprofen to treat arthritis face a 24 percent increased risk of suffering a heart attack." Feel the fear.

Almost everyone reported the relative risk increases: diclofenac increases the risk of heart attack by 55 percent; ibuprofen, by 24 percent. *The Boston Globe* was clever enough to report the natural frequencies: 1 extra heart attack in 1,005 people on ibuprofen. The U.K.'s *Daily Mirror*, meanwhile, tried and failed, reporting that 1 in 1,005 people on ibuprofen "will suffer heart failure over the following year." No. It's heart attack, not heart failure, and it's 1 *extra* person in 1,005, on top of the heart attacks you'd get anyway. Several other papers repeated the same mistake.

Often it's the fault of the press releases, and academics can themselves be as guilty as the rest when it comes to overdramatizing their research. But if anyone in a position of power is reading this, here is the information I would like from a newspaper, to help me make decisions about my health, when reporting on a risk: I want to know whom you're talking about (e.g., men in their fifties); I want to know what the baseline risk is (e.g., four men out of a hundred will have a heart attack over ten years); and I want to know what the increase in risk is, as a natural frequency (two extra men out of that hundred will have a heart attack over ten years). I also want to know exactly what's causing that increase in risk: an occasional headache pill, or a daily tubful of pain-relieving medication for arthritis. Then I will consider reading your newspapers again, instead of blogs that are written by people who understand research and that link reliably back to the original academic paper, so that I can double-check their précis when I wish.

More than a hundred years ago, H. G. Wells said that statistical thinking would one day be as important as the ability to read and write in a modern technological society. I disagree; probabilistic reasoning is difficult for everyone, but everyone understands normal numbers. This is why natural frequencies are the only sensible way to communicate risk.

CHOOSING YOUR FIGURES

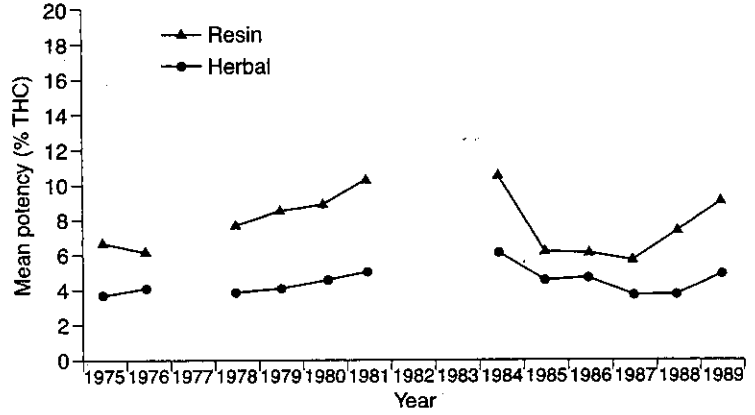
Sometimes the mispresentation of figures goes so far beyond reality that you can only assume mendacity. Often these situations seem to involve morality: drugs, abortion, and the rest. With very careful selection of numbers, in what some might consider to be a cynical and immoral manipulation of the facts for personal gain, you can sometimes make figures say anything you want.

The UK's *Independent* was in favor of legalizing cannabis for many years, but in March 2007 it decided to change its stance. One option would have been simply to explain this as a change of heart, or a reconsideration of the moral issues. Instead it was decorated with science—as cowardly zealots have done from eugenics through to prohibition—and justified with a fictitious change in the facts. CANNABIS—AN APOLOGY was the headline for its front-page splash: "In 1997, this newspaper launched a campaign to decriminalise the drug. If only we had known then what we can reveal today . . . Record numbers of teenagers are requiring drug treatment as a result of smoking skunk, the highly potent cannabis strain that is 25 times stronger than resin sold a decade ago." Twice in this story we are told that cannabis is twenty-five times stronger than it was a decade ago. For the paper's former editor Rosie Boycott, in her melodramatic recantation, skunk was "thirty times stronger." In one inside feature the strength issue was briefly downgraded to a "can be." The paper even referenced its figures: "The Forensic Science Service says that in the early Nineties cannabis would contain around 1 per cent tetrahydrocannabinol (THC), the mind-altering compound, but can now have up to 25 percent."

This is all sheer fantasy.

I've got the U.K.'s Forensic Science Service data right here in front of me, and the earlier data from the Laboratory of the Government Chemist, the United Nations Drug Control Program, and the European Monitoring Centre for Drugs and Drug Addiction. I'm going to share it with you, because I happen to think that people are very well able to make their own minds up about important social and moral issues when given the facts.

The data from the Laboratory of the Government Chemist goes from 1975 to 1989. Cannabis resin pootles around between 6 percent and 10 percent THC, herbal between 4 percent and 6 percent. There is no clear trend.



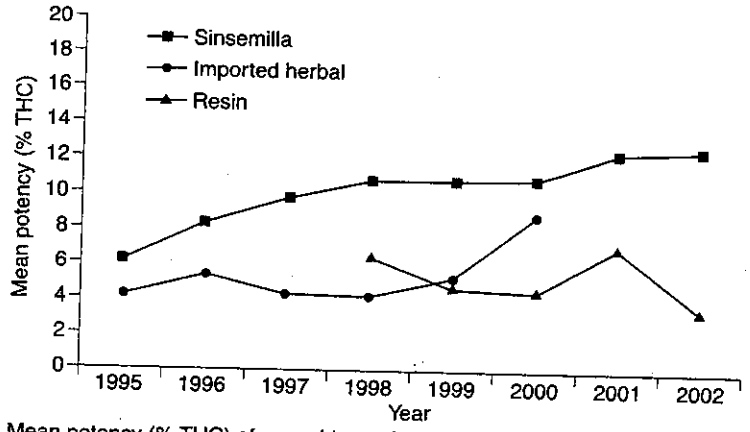
Mean potency (% THC) of cannabis products examined in the U.K. (Laboratory of the Government Chemist, 1975–89)

The Forensic Science Service data then takes over to produce the more modern figures, showing not much change in resin and domestically produced indoor herbal cannabis doubling in potency from 6 percent to around 12 or 14 percent (2003–05 data in table under references).

The rising trend of cannabis potency is gradual, fairly unspectacular, and driven largely by the increased availability of domestic, intensively grown indoor herbal cannabis.

"Twenty-five times stronger," remember. Repeatedly, and on the front page.

If you were in the mood to quibble with *The Independent*'s moral and political reasoning, as well as its evident and shameless venality, you could argue that intensive indoor cultivation of a plant that grows perfectly well outdoors is the cannabis industry's reaction to the product's illegality itself. It is dangerous to import cannabis in large amounts. It is dangerous to be caught growing a field of it. So it makes more sense to grow it intensively indoors, using expensive real estate, but producing a more concentrated drug. More concentrated drugs products are, after all, a natural conse-



Mean potency (% THC) of cannabis products examined in the U.K. (Forensic Science Service, 1995–2002)

Year	Sinsemilla %	Resin %	"Traditional" imported herbal %
1995	5.8	No data	3.9
1996	8.0	No data	5.0
1997	9.4	No data	4.0
1998	10.5	6.1	3.9
1999	10.6	4.4	5.0
2000	12.2	4.2	8.5
2001	12.3	6.7	No data
2002	12.3	3.2	No data
2003	12.0	4.6	No data
2004	12.7	1.6	No data
2005	14.2	6.6	No data

Mean THC content of cannabis products seized in the U.K. (Forensic Science Service, 1995–2005)

quence of illegality. You can't buy coca leaves in South London, but you can buy crack.

There is, of course, exceptionally strong cannabis to be found in some parts of the British market today, but then there always

has been. To get its scare figure, *The Independent* can only have compared the *worst* cannabis from the past with the *best* cannabis of today. It's an absurd thing to do, and moreover, you could have cooked the books in exactly the same way thirty years ago if you'd wanted; the figures for individual samples are available, and in 1975 the weakest herbal cannabis analyzed was 0.2 percent THC, while in 1978 the strongest herbal cannabis was 12 percent. By these figures, in just three years herbal cannabis became "sixty times stronger."

And this scare isn't even new. In the mid-1980s, during Ronald Reagan's "war on drugs," American campaigners were claiming that cannabis was fourteen times stronger than in 1970. Which sets you thinking. If it was fourteen times stronger in 1986 than in 1970, and it's twenty-five times stronger today than at the beginning of the 1990s, does that mean it's now 350 times stronger than in 1970?

That's not even a crystal in a plant pot. It's impossible. It would require more THC to be present in the plant than the total volume of space taken up by the plant itself. It would require matter to be condensed into superdense quark-gluon plasma cannabis. For God's sake don't tell the newspapers such a thing is possible.

COCAINE FLOODS THE PLAYGROUND

We are now ready to move on to some more interesting statistical issues, with another story from an emotive area, an article in *The Times* (London) in March 2006 headed: COCAINE FLOODS THE PLAYGROUND. "Use of the addictive drug by children doubles in a year," said the subheading. Was this true?

If you read the press release for the government survey on which the story is based, it reports "almost no change in patterns of drug use, drinking or smoking since 2000." But this was a gov-

ernment press release, and journalists are paid to investigate: perhaps the press release was hiding something, to cover up for government failures. The *Telegraph* also ran the "cocaine use doubles" story, and so did the *Mirror*. Did the journalists find the news themselves, buried in the report?

You can download the full document online. It's a survey of nine thousand children, aged eleven to fifteen, in 305 schools. The three-page summary said, again, that there was no change in prevalence of drug use. If you look at the full report, you will find the raw data tables: when asked whether they had used cocaine in the past year, 1 percent said yes in 2004, and 2 percent said yes in 2005.

So the newspapers were right: it doubled? No. Almost all the figures given were 1 percent or 2 percent. They'd all been rounded off. Civil servants are very helpful when you ring them up. The actual figures were 1.4 percent for 2004 and 1.9 percent for 2005, not 1 percent and 2 percent. So cocaine use hadn't doubled at all. But people were still eager to defend this story; cocaine use, after all, had increased, yes?

No. What we now have is a relative risk increase of 35.7 percent, or an absolute risk increase of 0.5 percent. If we use the real numbers, out of nine thousand kids we have about forty-five more saying yes to the question "Did you take cocaine in the past year?"

Presented with a small increase like this, you have to think: Is it statistically significant? I did the math, and the answer is yes, it is, in that you get a p-value of less than 0.05. What does "statistically significant" mean? It's just a way of expressing the likelihood that the result you got was attributable merely to chance. Sometimes you might throw heads five times in a row, with a completely normal coin, especially if you kept tossing it for long enough. Imagine a jar of 980 blue marbles, and 20 red ones, all mixed up; every now and then—albeit rarely—picking blindfolded, you might pull out 3 red ones in a row, just by chance. The standard cutoff point for statistical significance is a p-value of 0.05, which is just

another way of saying, "If I did this experiment a hundred times, I'd expect a spurious positive result on five occasions, just by chance."

To go back to our concrete example of the kids in the playground, let's imagine that there was definitely no difference in cocaine use, but you conducted the same survey a hundred times. You might get a difference like the one we have seen here, just by chance, just because you randomly happened to pick up more of the kids who had taken cocaine this time around. But you would expect this to happen less than five times out of your hundred surveys.

So we have a risk increase of 35.7 percent, which seems at face value to be statistically significant; but it is an isolated figure. To "data mine," taking it out of its real-world context and saying it is significant, is misleading. The statistical test for significance assumes that every data point is independent, but here the data is "clustered," as statisticians say. They are not data points; they are real children, in 305 schools. They hang out together; they copy one another; they buy drugs from one another; there are crazes, epidemics, group interactions.

The increase of forty-five kids taking cocaine could have been a massive epidemic of cocaine use in one school, or a few groups of a dozen kids in a few different schools, or miniepidemics in a handful of schools. Or forty-five kids independently sourcing and consuming cocaine alone without their friends, which seems pretty unlikely to me.

This immediately makes our increase less statistically significant. The small increase of 0.5 percent was significant only because it came from a large sample of nine thousand data points—like nine thousand tosses of a coin—and the one thing almost everyone knows about studies like this is that a bigger sample size means the results are probably more significant. But if they're not independent data points, then you have to treat it, in some respects, like a smaller sample, so the results become less significant. As stat-

isticians would say, you must "correct for clustering." This is done with clever math that makes everyone's head hurt. All you need to know is that the reasons why you must correct for clustering are transparent, obvious, and easy, as we have just seen (in fact, as with many implements, knowing when to use a statistical tool is a different and equally important skill from understanding how it is built). When you correct for clustering, you greatly reduce the significance of the results. Will our increase in cocaine use, already down from "doubled" to "35.7 percent," even survive?

No. Because there is a final problem with this data: there is so much of it to choose from. There are dozens of data points in the report: on solvents, cigarettes, ketamine, cannabis, and so on. It is standard practice in research that we only accept a finding as significant if it has a p-value of 0.05 or less. But as we said, a p-value of 0.05 means that for every hundred comparisons you do, five will be positive by chance alone. From this report you could have done dozens of comparisons, and some of them would indeed have shown increases in usage—but by chance alone, and the cocaine figure could be one of those. If you roll a pair of dice often enough, you will get a double six three times in a row on many occasions. This is why statisticians do a "correction for multiple comparisons," a correction for "rolling the dice" lots of times. This, like correcting for clustering, is particularly brutal on the data and often reduces the significance of findings dramatically.

Data dredging is a dangerous profession. You could—at face value, knowing nothing about how stats works—have said that this government report showed a significant increase of 35.7 percent in cocaine use. But the stats nerds who compiled it knew about clustering and Bonferroni's correction for multiple comparisons. They are not stupid; they do stats for a living.

That, presumably, is why they said quite clearly in their summary, in their press release, and in the full report that there was no change from 2004 to 2005. But the journalists did not want to

believe this. They tried to reinterpret the data for themselves; they looked under the hood, and they thought they'd found the news. The increase went from 0.5 percent—a figure that might be a gradual trend, but could equally well be an entirely chance finding—to a front-page story in *The Times* (London) about cocaine use's doubling. You might not trust the press release, but if you don't know about numbers, then you take a big chance when you delve under the hood of a study to find a story.

OK, BACK TO AN EASY ONE

There are also some perfectly simple ways to generate ridiculous statistics, and two common favorites are to select an unusual sample group of people and to ask them a stupid question. Let's say 70 percent of all women want Prince Charles to be told to stop interfering in public life. Oh, hang on—70 percent of all women *who visit my website* want Prince Charles to be told to stop interfering in public life. You can see where we're going. And of course, in surveys, if they are voluntary, there is something called selection bias; only the people who can be bothered to fill out the survey form will actually have a vote registered.

There was an excellent example of this in *The Daily Telegraph* in the last days of 2007. DOCTORS SAY NO TO ABORTIONS IN THEIR SURGERIES was the headline. "Family doctors are threatening a revolt against government plans to allow them to perform abortions in their surgeries, the *Daily Telegraph* can disclose." A revolt? "Four out of five doctors do not want to carry out terminations even though the idea is being tested in NHS [National Health Service] pilot schemes, a survey has revealed."

Where did these figures come from? A systematic survey of all doctors, with lots of chasing to catch the nonresponders? Telephoning them at work? A postal survey, at least? No. It was an

online vote on a doctors' chat site that produced this major news story. Here is the question and the options given:

> "[Doctors] should carry out abortions in their surgeries"
>
> Strongly agree, agree, don't know, disagree, strongly disagree.

We should be clear: I myself do not fully understand this question. Is that "should" as in "should"? As in "ought to"? And in what circumstances? With extra training, time, and money? With extra systems in place for adverse outcomes? And remember, this is a website where doctors—bless them—go to moan. Are they just saying no because they're grumbling about more work and low morale?

More than that, what exactly does "abortion" mean here? Looking at the comments in the chat forum, I can tell you that plenty of the doctors seemed to think it was about surgical abortions, not the relatively safe oral pill for termination of pregnancy. Doctors aren't that bright, you see. Here are some quotes:

> This is a preposterous idea. How can [doctors] ever carry out abortions in their own surgeries. What if there was a major complication like uterine and bowel perforation?

> [Doctor's] surgeries are the places par excellence where infective disorders present. The idea of undertaking there any sort of sterile procedure involving an abdominal organ is anathema.

> The only way it would or rather should happen is if [doctor] practices have a surgical day care facility as part of their premises which is staffed by appropriately trained staff, i.e. theater staff, anesthetist and gynecologist . . . any surgical operation is not without its risks, and presumably

> [we] will undergo gynecological surgical training in order to perform.

> What are we all going on about? Let's all carry out abortions in our surgeries, living rooms, kitchens, garages, corner shops, you know, just like in the old days.

And here's my favorite:

> I think that the question is poorly worded and I hope that [the doctors' website] do[es] not release the results of this poll to *The Daily Telegraph*.

BEATING YOU UP

It would be wrong to assume that the kinds of oversights we've covered so far are limited to the lower echelons of society, like doctors and journalists. Some of the most sobering examples come from the very top.

In 2006, after a major British government report, the media reported that one murder a week is committed by someone with psychiatric problems. Psychiatrists should do better, the newspapers told us, and prevent more of these murders. All of us would agree, I'm sure, with any sensible measure to improve risk management and violence, and it's always timely to have a public debate about the ethics of detaining psychiatric patients (although in the name of fairness I'd like to see preventive detention discussed for all other potentially risky groups too—like alcoholics, the repeatedly violent, people who have abused staff in the job center, and so on).

But to engage in this discussion, you need to understand the math of predicting very rare events. Let's take a very concrete example and look at the HIV test. What features of any diagnostic procedure do we measure in order to judge how useful it might be?

Statisticians would say the blood test for HIV has a very high "sensitivity," at 0.999. That means that if you do have the virus, there is a 99.9 percent chance that the blood test will be positive. They would also say the test has a high "specificity" of 0.9999, so if you are not infected, there is a 99.99 percent chance that the test will be negative. What a smashing blood test.*

But if you look at it from the perspective of the person being tested, the math gets slightly counterintuitive. Because weirdly, the meaning, the predictive value, of an individual's positive or negative test is changed in different situations, depending on the background rarity of the event that the test is trying to detect. The rarer the event in your population, the worse your test becomes, even though it is the same test.

This is easier to understand with concrete figures. Let's say the HIV infection rate among high-risk men in a particular area is 1.5 percent. We use our excellent blood test on 10,000 of these men, and we can expect 151 positive blood results overall: 150 will be our truly HIV positive men, who will get true positive blood tests, and 1 will be the one false positive we could expect from having 10,000 HIV negative men being given a test that is wrong one time in 10,000. So, if you get a positive HIV blood test result, in these circumstances your chances of being truly HIV positive are 150 out of 151. It's a highly predictive test.

Let's now use the same test where the background HIV infection rate in the population is about one in ten thousand. If we test ten thousand people, we can expect two positive blood results overall: one from the person who really is HIV positive; and the one false positive that we could expect, again, from having ten thousand HIV negative men being tested with a test that is wrong one time in ten thousand.

Suddenly, when the background rate of an event is rare, even our previously brilliant blood test becomes a bit rubbish. For the

*The figures here are ballpark, from Gerd Gigerenzer's excellent book *Reckoning with Risk*.

two men with a positive HIV blood test result, in this population where only one in ten thousand has HIV, it's only fifty-fifty odds on whether they really are HIV positive.

Let's think about violence. The best predictive tool for psychiatric violence has a sensitivity of 0.75 and a specificity of 0.75. It's tougher to be accurate when we predict an event in humans, with human minds and changing human lives. Let's say 5 percent of patients seen by a community mental health team will be involved in a violent event in a year. If we use the same math as we did for the HIV tests, your "0.75" predictive tool would be wrong eighty-six times out of a hundred. For serious violence, occurring at 1 percent a year, with our best "0.75" tool, you inaccurately finger your potential perpetrator ninety-seven times out of a hundred. Will you preventively detain ninety-seven people to prevent three violent events? And will you apply that rule to alcoholics and assorted nasty antisocial types as well?

For murder, the extremely rare crime in question in this report, for which more action was demanded, occurring at one in ten thousand a year among patients with psychosis, the false positive rate is so high that the best predictive test is entirely useless.

This is not a counsel of despair. There are things that can be done, and you can always try to reduce the number of actual stark cock-ups, although it's difficult to know what proportion of the "one murder a week" represents a clear failure of a system, since when you look back in history, through the retrospectoscope, anything that happens will look as if it were inexorably leading up to your one bad event. I'm just giving you the math on rare events. What you do with it is a matter for you.

LOCKING YOU UP

In 1999 British lawyer Sally Clark was put on trial for murdering her two babies. In the U.K. this was a major trial, with a successful

appeal, and although many have a dim awareness that there was a statistical error in the prosecution case, few know the true story or the phenomenal extent of the statistical ignorance that went on in the case.

At her trial, Professor Sir Roy Meadow, an expert in parents who harm their children, was called to give expert evidence. Meadow famously quoted "one in seventy-three million" as the chance of two children in the same family dying of sudden infant death syndrome (SIDS).

This was a very problematic piece of evidence for two very distinct reasons: one is easy to understand; the other is an absolute mind bender. Because you have the concentration span to follow the next two pages, you will come out smarter than Professor Sir Roy, the judge in the Sally Clark case, her defense teams, the appeal court judges, and almost all the journalists and legal commentators reporting on the case. We'll do the easy reason first.

THE ECOLOGICAL FALLACY

The figure of "one in seventy-three million" itself is iffy, as everyone now accepts. It was calculated as 8,543 × 8,543, as if the chances of two SIDS episodes in this one family were independent of each other. This feels wrong from the outset, and anyone can see why: there might be environmental or genetic factors at play, both of which would be shared by the two babies. But forget how pleased you are with yourself for understanding that fact. Even if we accept that two SIDS in one family is much more likely than one in seventy-three million—say, one in ten thousand—any such figure is still of dubious relevance, as we shall now see.

THE PROSECUTOR'S FALLACY

The real question in this case is: What do we do with this spurious number? Many press reports at the time stated that one in seventy-three million was the likelihood that the deaths of Sally Clark's two children were accidental—that is, the likelihood that she was innocent. Many in the court process seemed to share this view, and the factoid certainly sticks in the mind. But this is an example of a well-known and well-documented piece of flawed reasoning known as the prosecutor's fallacy.

Two babies in one family have died. This in itself is very rare. Once this rare event has occurred, the jury needs to weigh up two competing explanations for the babies' deaths: double SIDS or double murder. Under normal circumstances—before any babies have died—double SIDS is very unlikely, and so is double murder. But now that the rare event of two babies dying in one family has occurred, the two explanations—double murder or double SIDS—are suddenly both very likely. If we really wanted to play statistics, we would need to know which is relatively *more* rare: double SIDS or double murder. People have tried to calculate the relative risks of these two events, and one paper says it comes out at around two to one in favor of double SIDS.

Not only was this *crucial* nuance of the prosecutor's fallacy missed at the time—by everyone in the court—but it was also clearly missed in the appeal, at which the judges suggested that instead of "one in seventy-three million," Meadow should have said "very rare." They recognized the flaws in its calculation, the ecological fallacy, the easy problem above, but they still accepted his number as establishing "a very broad point, namely the rarity of double SIDS."

That, as you now understand, was entirely wrongheaded; the rarity of double SIDS is irrelevant, because double murder is rare too. An entire court process failed to spot the nuance of how the figure should be used. Twice.

Meadow was foolish, and has been vilified (some might say this process was exacerbated by the witch hunt against pediatricians who work on child abuse), but if it is true that he should have spotted and anticipated the problems in the interpretation of

his number, then so should the rest of the people involved in the case: a pediatrician has no more unique responsibility to be numerate than a lawyer, a judge, journalist, jury member, or clerk. The prosecutor's fallacy is also highly relevant in DNA evidence, for example, in which interpretation frequently turns on complex mathematical and contextual issues. Anyone who is going to trade in numbers, and use them, and think with them, and persuade with them, let alone lock people up with them, also has a responsibility to understand them. All you've done is read a popular science book on them, and already you can see it's hardly rocket science.

LOSING THE LOTTERY

> You know, the most amazing thing happened to me tonight. I was coming here, on the way to the lecture, and I came in through the parking lot. And you won't believe what happened. I saw a car with the license plate ARW 357. Can you imagine? Of all the millions of license plates in the state, what was the chance that I would see that particular one tonight? Amazing . . . —Richard Feynman

It is possible to be very unlucky indeed. A nurse named Lucia de Berk has been in prison for six years in Holland, convicted of seven counts of murder and three of attempted murder. An unusually large number of people died when she was on shift, and that, essentially, along with some very weak circumstantial evidence, is the substance of the case against her. She has never confessed, she has continued to protest her innocence, and her trial has generated a small collection of theoretical papers in the statistics literature.

The judgment was largely based on a figure of "one in 342 million against." Even if we found errors in this figure—and believe me, we will—as in our previous story, the figure itself would still be largely irrelevant. Because, as we have already seen repeatedly,

the interesting thing about statistics is not the tricky math, but what the numbers mean.

There is also an important lesson here from which we could all benefit: unlikely things do happen. Somebody wins the lottery every week; children are struck by lightning. It's only weird and startling when something very, very specific and unlikely happens if you have specifically predicted it beforehand.*

Here is an analogy.

Imagine I am standing near a large wooden barn with an enormous machine gun. I place a blindfold over my eyes, and laughing maniacally, I fire off many thousands and thousands of bullets into the side of the barn. I then drop the gun, walk over to the wall, examine it closely for some time, all over, pacing up and down. I find one spot where there are three bullet holes close to one another, then draw a target around them, announcing proudly that I am an excellent marksman.

You would, I think, disagree with both my methods and my conclusions for that deduction. But this is exactly what has happened in Lucia's case: the prosecutors found seven deaths on one nurse's shifts, in one hospital, in one city, in one country, in the world and then drew a target around them.

This breaks a cardinal rule of any research involving statistics: you cannot find your hypothesis in your results. Before you go to your data with your statistical tool, you have to have a specific hypothesis to test. If your hypothesis comes from analyzing the data, then there is no sense in analyzing the same data again to confirm it.

This is a rather complex, philosophical, mathematical form of circularity, but there were also very concrete forms of circular rea-

*The magician and pseudoscience debunker James Randi used to wake up every morning and write on a card in his pocket: "I, James Randi, will die today," followed by the date and his signature. Just in case, he has recently explained, he really did, by some completely unpredictable accident.

soning in the case. To collect more data, the investigators went back to the wards to see if they could find more suspicious deaths. But all the people who were asked to remember "suspicious incidents" knew that they were being asked because Lucia might be a serial killer. There was a high risk that "an incident was suspicious" became synonymous with "Lucia was present." Some sudden deaths when Lucia was not present would not be listed in the calculations, by definition: they are in no way suspicious, because Lucia was not present.

It gets worse. "We were asked to make a list of incidents that happened during or shortly after Lucia's shifts," said one hospital employee. In this manner more patterns were unearthed, and so it became even more likely that investigators would find more suspicious deaths on Lucia's shifts. Meanwhile, Lucia waited in prison for her trial.

This is the stuff of nightmares.

At the same time, a huge amount of corollary statistical information was almost completely ignored. In the three years before Lucia worked on the ward in question, there were seven deaths. In the three years that she did work on the ward, there were six deaths. Here's a thought: it seems odd that the death rate should go *down* on a ward at the precise moment that a serial killer—on a killing spree—arrives. If Lucia killed them all, then there must have been no natural deaths on that ward at all in the whole of the three years that she worked there.

Ah, but on the other hand, as the prosecution revealed at her trial, Lucia did like tarot. And she does sound a bit weird in her private diary, excerpts from which were read out. So she might have done it anyway.

But the strangest thing of all is this. In generating his obligatory, spurious, Meadowesque figure, which this time was "one in 342 million," the prosecution's statistician made a simple, rudimentary mathematical error. He combined individual statistical tests

by multiplying p-values, the mathematical description of chance, or statistical significance. This bit's for the hard-core science nerds, and will be edited out by the publisher, but I intend to write it anyway: you do not just multiply p-values together; you weave them with a clever tool, like maybe "Fisher's method for combination of independent p-values."

If you multiply p-values together, then harmless and probable incidents rapidly appear vanishingly unlikely. Let's say you worked in twenty hospitals, each with a harmless incident pattern, say, $p=0.5$. If you multiply those harmless p-values, of entirely chance findings, you end up with a final p-value of 0.5 to the power of twenty, which is $p < 0.000001$, which is extremely, very, highly statistically significant. With this mathematical error, by his reasoning, if you change hospitals a lot, you automatically become a suspect. Have you worked in twenty hospitals? For God's sake, don't tell the Dutch police if you have.