THE NEW YORKER

WHAT STATISTICS CAN AND CAN'T TELL US ABOUT OURSELVES

In the era of Big Data, we've come to believe that, with enough information, human behavior is predictable. But number crunching can lead us perilously wrong.

By Hannah Fry September 2, 2019

Making individual predictions from collective characteristics is a risky business.
Illustration by Ben Wiseman

0:00 / 23:27

Audio: Listen to this article. To hear more, download Audm for iPhone or Android.

Harold Eddleston, a seventy-seven-year-old from Greater Manchester, was still reeling from a cancer diagnosis he had been given that week when, on a Saturday morning in February, 1998, he received the worst possible news. He would have to face the future alone: his beloved wife had died unexpectedly, from a heart attack.

Eddleston's daughter, concerned for his health, called their family doctor, a well-respected local man named Harold Shipman. He came to the house, sat with her father, held his hand, and spoke to him tenderly. Pushed for a prognosis as he left, Shipman replied portentously, "I wouldn't buy him any Easter eggs." By Wednesday, Eddleston was dead; Dr. Shipman had murdered him.

Harold Shipman was one of the most prolific serial killers in history. In a twenty-three-year career as a mild-mannered and well-liked family doctor, he injected at least two hundred and fifteen of his patients with lethal doses of opiates. He was finally arrested in September, 1998, six months after Eddleston's death.

David Spiegelhalter, the author of an important and comprehensive new

book, "The Art of Statistics" (Basic), was one of the statisticians tasked by the ensuing public inquiry to establish whether the mortality rate of Shipman's patients should have aroused suspicion earlier. Then a biostatistician at Cambridge, Spiegelhalter found that Shipman's excess mortality—the number of his older patients who had died in the course of his career over the number that would be expected of an average doctor's—was a hundred and seventy-four women and forty-nine men at the time of his arrest. The total closely matched the number of victims confirmed by the inquiry.

One person's actions, written only in numbers, tell a profound story. They gesture toward the unimaginable grief caused by one man. But at what point do many deaths become too many deaths? How do you distinguish a suspicious anomaly from a run of bad luck? For that matter, how can we know in advance the number of people we expect to die? Each death is preceded by individual circumstances, private stories, and myriad reasons; what does it mean to wrap up all that uncertainty into a single number?

In 1825, the French Ministry of Justice ordered the creation of a national collection of crime records. It seems to have been the first of its kind anywhere in the world—the statistics of every arrest and conviction in the country, broken down by region, assembled and ready for analysis. It's the kind of data set we take for granted now, but at the time it was extraordinarily novel. This was an early instance of Big Data—the first time that mathematical analysis had been applied in earnest to the messy and unpredictable realm of human behavior.

Or maybe not so unpredictable. In the early eighteen-thirties, a Belgian astronomer and mathematician named Adolphe Quetelet analyzed the numbers and discovered a remarkable pattern. The crime records were startlingly consistent. Year after year, irrespective of the actions of courts and prisons, the number of murders, rapes, and robberies reached almost exactly the same total. There is a "terrifying exactitude with which crimes reproduce

themselves," Quetelet said. "We know in advance how many individuals will dirty their hands with the blood of others. How many will be forgers, how many poisoners."

To Quetelet, the evidence suggested that there was something deeper to discover. He developed the idea of a "Social Physics," and began to explore the possibility that human lives, like planets, had an underlying mechanistic trajectory. There's something unsettling in the idea that, amid the vagaries of choice, chance, and circumstance, mathematics can tell us something about what it is to be human. Yet Quetelet's overarching findings still stand: at some level, human life can be quantified and predicted. We can now forecast, with remarkable accuracy, the number of women in Germany who will choose to have a baby each year, the number of car accidents in Canada, the number of plane crashes across the Southern Hemisphere, even the number of people who will visit a New York City emergency room on a Friday evening.

In some ways, this is what you would expect from any large, disordered

system. Think about the predictable and quantifiable way that gases behave. It might be impossible to trace the movement of each individual gas molecule, but the uncertainty and disorder at the molecular level wash out when you look at the bigger picture. Similarly, larger regularities emerge from our individually unpredictable lives. It's almost as though we woke up each morning with a chance, that day, of becoming a murderer, causing a car accident, deciding to propose to our partner, being fired from our job. "An assumption of 'chance' encapsulates all the inevitable unpredictability in the world," Spiegelhalter writes.

But it's one thing when your aim is to speak in general terms about who we are together, as a collective entity. The trouble comes when you try to go the other way—to learn something about us as individuals from how we behave as a collective. And, of course, those answers are often the ones we most want.

The dangers of making individual predictions from our collective characteristics were aptly demonstrated in a deal struck by the French lawyer André-François Raffray in 1965. He agreed to pay a ninety-year-old woman twenty-five hundred francs every month until her death, whereupon he would take possession of her apartment in Arles.

At the time, the average life expectancy of French women was 74.5 years, and Raffray, then forty-seven, no doubt thought he'd negotiated himself an auspicious contract. Unluckily for him, as Bill Bryson recounts in his new book, "The Body," the woman was Jeanne Calment, who went on to become the oldest person on record. She survived for thirty-two years after their deal was signed, outliving Raffray, who died at seventy-seven. By then, he had paid more than twice the market value for an apartment he would never live in.

Raffray learned the hard way that people are not well represented by the average. As the mathematician Ian Stewart points out in "Do Dice Play

God?" (Basic), the average person has one breast and one testicle. In large groups, the natural variability among human beings cancels out, the random zig being countered by the random zag; but that variability means that we can't speak with certainty about the individual—a fact with wide-ranging consequences.

Every day, millions of people, David Spiegelhalter included, swallow a small white statin pill to reduce the risk of heart attack and stroke. If you are one of those people, and go on to live a long and happy life without ever suffering a heart attack, you have no way of knowing whether your daily statin was responsible or whether you were never going to have a heart attack in the first place. Of a thousand people who take statins for five years, the drugs will help only eighteen to avoid a major heart attack or stroke. And if you do find yourself having a heart attack you'll never know whether it was delayed by taking the statin. "All I can ever know," Spiegelhalter writes, "is that on average it benefits a large group of people like me."

That's the rule with preventive drugs: for most individuals, most of those drugs won't do anything. The fact that they produce a collective benefit makes them worth taking. But it's a pharmaceutical form of Pascal's wager: you may as well act as though God were real (and believe that the drugs will work for you), because the consequences otherwise outweigh the inconvenience.

There is so much that, on an individual level, we don't know: why some people can smoke and avoid lung cancer; why one identical twin will remain healthy while the other develops a disease like A.L.S.; why some otherwise similar children flourish at school while others flounder. Despite the grand promises of Big Data, uncertainty remains so abundant that specific human lives remain boundlessly unpredictable. Perhaps the most successful prediction engine of the Big Data era, at least in financial terms, is the Amazon recommendation algorithm. It's a gigantic statistical machine worth a huge sum to the company. Also, it's wrong most of the time. "There is

nothing of chance or doubt in the course before my son," Dickens's Mr. Dombey says, already imagining the business career that young Paul will enjoy. "His way in life was clear and prepared, and marked out before he existed." Paul, alas, dies at age six.

And yet, amid the oceans of unpredictability, we've somehow managed not to drown. Statisticians have navigated a route to maximum certainty in an uncertain world. We might not be able to address insular quandaries, like "How long will I live?," but questions like "How many patient deaths are too many?" can be tackled. In the process, a powerful idea has arisen to form the basis of modern scientific research.

A stranger hands you a coin. You have your suspicions that it's been weighted somehow, perhaps to make heads come up more often. But for now you'll happily go along with the assumption that the coin is fair.

MORE FROM THIS ISSUE
SEPTEMBER 9, 2019

PORTFOLIO
Alex Prager's L.A. Dreaming
By Alex Prager

BOOKS
Briefly Noted Book Reviews

THE SHARING
At Last, a Pools
By Bruce H

You toss the coin twice, and get two heads in a row. Nothing to get excited about just yet. A perfectly fair coin will throw two heads in a row twenty-five per cent of the time—a probability known as the p-value. You keep tossing and get another head. Then another. Things are starting to look fishy, but

even if you threw the coin a thousand times, or a million, you could never be absolutely sure it was rigged. The chances might be minuscule, but in theory a fair coin could still produce any combination of heads.

Scientists have picked a path through all this uncertainty by setting an arbitrary threshold, and agreeing that anything beyond that point gives you grounds for suspicion. Since 1925, when the British statistician Ronald Fisher first suggested the convention, that threshold has typically been set at five per cent. You're seeing a suspicious number of heads, and once the chance of a fair coin turning up at least as many heads as you've seen dips below five per cent, you can abandon your stance of innocent until proved guilty. In this case, five heads in a row, with a p-value of 3.125 per cent, would do it.

This is the underlying principle behind how modern science comes to its conclusions. It doesn't matter if we're uncovering evidence for climate change or deciding whether a drug has an effect: the concept is identical. If the results are too unusual to have happened by chance—at least, not more than one time out of twenty—you have reason to think that your hypothesis has been vindicated. "Statistical significance" has been established.

Take a clinical trial on aspirin run by the Oxford medical epidemiologist Richard Peto in 1988. Aspirin interferes with the formation of blood clots, and can be used to prevent them in the arteries of the heart or the brain. Peto's team wanted to know whether aspirin increased your chances of survival if it was administered in the middle of a heart attack.

Their trial involved 17,187 people and showed a remarkable effect. In the group that was given a placebo, 1,016 patients died; of those who had taken the aspirin, only 804 died. Aspirin didn't work for everyone, but it was unlikely that so many people would have survived if the drug did nothing. The numbers passed the threshold; the team concluded that the aspirin was working.

Such statistical methods have become the currency of modern research. They've helped us to make great strides forward, to find signals in noisy data. But, unless you are extraordinarily careful, trying to erase uncertainty comes with downsides. Peto's team submitted the results of their experiment to an illustrious medical journal, which came back with a request from a referee: could Peto and his colleagues break the results down into groups? The referee wanted to know how many women had been saved by the aspirin, how many men, how many with diabetes, how many in this or that age bracket, and so on.

Peto objected. By subdividing the big picture, he argued, you introduce all kinds of uncertainty into the results. For one thing, the smaller the size of the groups considered, the greater the chance of a fluke. It would be "scientifically stupid," he observed, to draw conclusions on anything other than the big picture. The journal was insistent, so Peto relented. He resubmitted the paper with all the subgroups the referee had asked for, but with a sly addition. He also subdivided the results by astrological sign. It wasn't that astrology was going to influence the impact of aspirin; it was that, just by chance, the number of people for whom aspirin works will be greater in some groups than in others. Sure enough, in the study, it appeared as though aspirin didn't work for Libras and Geminis but halved your risk of death if you happened to be a Capricorn.

Using sufficiently large groups might help to insure against flukes, but there's another trap that befalls unsuspecting scientists. It's one that Peto's experiment also serves to underline, and one that has led to nothing less than a statistical crisis at the heart of science.

The easiest way to understand the issue is by returning to the conundrum of the biased coin. (Coins are the statistician's pet example for a reason.) Suppose that you're particularly keen not to draw a false conclusion, and decide to hang on to your hypothesis that the coin is fair unless you get

twenty heads in a row. A fair coin would do this only about one in a million times, so it's an extraordinarily high level of proof to demand—far beyond the threshold of five per cent used by much of science.

Now, imagine I gave out fair coins to every person in the United States and asked everyone to complete the same test. Here's the issue: even with a threshold of one in a million—even with everything perfectly fair and aboveboard—we would still expect around three hundred of these people to throw twenty heads in a row. If they were following Fisher's method, they'd have no choice but to conclude that they'd been given a trick coin. The fact is that, wherever you decide to set the threshold, if you repeat your experiment enough times, extremely unlikely outcomes are bound to arise eventually.

Apple learned this shortly after the iPod Shuffle was launched. The device would play songs from a users' library at random, but Apple found itself inundated with complaints from users who were convinced that their Shuffle was playing songs in a pattern. Patterns are much more likely to occur than we think, but even if several songs by the same artist, or consecutive songs from an album, had only a tiny probability of appearing next to one another in the playlist, so many people were listening to their iPods that it was inevitable such seemingly strange coincidences would occur.

In science, the situation is starker, and the stakes are higher. With a threshold of only five per cent, one in twenty studies will inadvertently find evidence for nonexistent phenomena in its data. That's another reason that Peto resisted the proposal that he look at various subpopulations: the greater the number of groups you look at, the greater your chances of seeing spurious effects. And this is far from being only a theoretical concern. In medicine, a study of forty-nine of the most cited medical publications from 1990 to 2003 found that the conclusions of sixteen per cent were contradicted by subsequent studies. Psychology fares worse still in these surveys (possibly because its studies are cheaper to reproduce). A 2015 study found that attempts to reproduce a

hundred psychological experiments yielded significant results in only thirty-six per cent of them, even though ninety-seven per cent of the initial studies reported a p-value under the five-per-cent threshold. And scientists fear that, as with the iPod Shuffle, the fluke results tend to get an outsized share of attention.

Many high-profile studies are now widely believed to have been founded on such flukes. You may have come across the research on power posing, which suggests that adopting a dominant stance helps to reduce stress hormones in the body. The study has a thousand citations, and an accompanying TED talk has amassed more than fifty million views, but the findings have failed to be replicated and are now regarded as a notable example of the flaws in Fisher's methods.

It's not that scientific fraud is common; it's that too many researchers have failed to handle uncertainty with sufficient care. This issue has only been exacerbated in the era of Big Data. The more data that are collected, cross-referenced, and searched for correlations, the easier it becomes to reach false conclusions. Illustrating this point, Spiegelhalter includes a 2009 study in which researchers put a subject into an fMRI scanner and analyzed the response in 8,064 brain sites while showing photographs of different human expressions. The scientists wanted to see what regions of the brain were lighting up in response to the photographs and used a threshold of a tenth of one per cent for their experiment. "The twist was that the 'subject' was a 4lb Atlantic Salmon which 'was not alive at the time of scanning,' " Spiegelhalter notes.

But, even at that threshold, run enough tests and you're bound to cross it eventually. Of the more than eight thousand sites in the dead fish's brain the researchers inspected, sixteen duly showed a statistically significant response. And the fear is that equally unfounded conclusions, albeit less apparently so, will routinely be drawn, with the false assurance of "statistical significance."

Science still stands up to scrutiny, precisely because it invites scrutiny. But the p-value crisis suggests that our current procedures could be improved upon.

Scientists now say that researchers should declare their hypothesis in advance of a study, in order to make fishing for significant results much more difficult. Most agree that the incentives of science need to be changed, too—that studies designed to replicate the work of others should be valued more highly. There are also suggestions for an alternative way to present experimental findings. Many people have called for the focus of science to be on the size of the effect—how many lives are saved by a drug, for instance—rather than on whether the data for some effect cross some arbitrary threshold. How impressed should we be by very strong evidence for a very weak effect? Let's go back to aspirin. A gigantic study—it tracked twenty-two thousand individuals over five years—demonstrated that taking small daily doses of the drug would reduce the risk of a heart attack. The p-value, the probability of this (or something more extreme) happening by chance, was tiny: 0.001 per cent. But so, too, was the effect size. A hundred and thirty otherwise healthy individuals would have to take the drug to prevent a single heart attack, and all the while each person would be increasing his or her risk of adverse side effects. It's a risk that is now deemed to outweigh the benefits for most people, and the advice for older adults to take a baby aspirin a day has recently been recanted.

But perhaps the real problem is how difficult we find it to embrace uncertainty. Earlier this year, eight hundred and fifty prominent academics, including David Spiegelhalter, signed a letter to *Nature* arguing that the issue can't be solved with a technical work-around. P-values aren't the problem; the problem is our obsession with setting a threshold.

Drawing an arbitrary line in the sand creates an illusion that we can divide the true from the false. But the results of a complicated experiment cannot be reduced to a yes-or-no answer. Back when Spiegelhalter was asked to

determine whether Dr. Harold Shipman's mortality rate should have aroused suspicion earlier, he swiftly decided that the standard test of statistical significance would be a "grossly inappropriate" way to monitor doctors. The medical profession would effectively be pointing the finger of suspicion at one in every twenty innocent doctors—thousands of clinicians in the U.K. Doctors would be penalized for treating higher-risk patients.

Instead, Spiegelhalter and his colleagues proposed an alternative test, which took account of patient deaths as they occurred, contrasting the accumulating deaths with the expected number. Year on year, it sequentially compares the likelihood that a doctor's high mortality rates are a run of bad luck with something more suspicious, and raises an alarm once the evidence starts to build. But even this highly sophisticated method will, owing to the capricious whims of chance, eventually cast suspicion on the innocent. Indeed, as soon as a monitoring system for general practitioners was piloted, it "immediately identified a G.P. with even higher mortality rates than Shipman," Spiegelhalter writes. This was an unlucky doctor who worked in a coastal town with an elderly population. The result highlights how careful you need to be even with the best statistical methods. In Spiegelhalter's words, while statistics can find the outliers, it "cannot offer reasons why these might have occurred, so they need careful implementation in order to avoid false accusations."

Statistics, for all its limitations, has a profound role to play in the social realm. The Shipman inquiry concluded that, if such a monitoring system had been in place, it would have raised the alarm as early as 1984. Around a hundred and seventy-five lives could have been saved. A mathematical analysis of what it is to be human can take us only so far, and, in a world of uncertainty, statistics will never eradicate doubt. But one thing is for sure: it's a very good place to start. ♦

An earlier version of this article incorrectly defined p-value.

This article appears in the print edition of the September 9, 2019, issue, with the headline "Your Number's Up."

Hannah Fry is a professor at University College London's Centre for Advanced Spatial Analysis. Her latest book is "Hello World." Read more »

Video

Why the Amazon Fires Are Surging
Wildfires have long occurred in the Amazon rain forest, but never on this scale. The New Yorker staff writer Jon Lee Anderson explains how they began, and what will happen if the planet's great green lung continues to burn.