The Law of Small Numbers

A study of the incidence of kidney cancer in the 3,141 counties of the United a>< НЉStates reveals a remarkable pattern. The counties in which the incidence of kidney cancer is lowest are mostly rural, sparsely populated, and located in traditionally Republican states in the Midwest, the South, and the West. What do you make of this?

Your mind has been very active in the last few seconds, and it was mainly a System 2 operation. You deliberately searched memory and formulated hypotheses. Some effort was involved; your pupils dilated, and your heart rate increased measurably. But System 1 was not idle: the operation of System 2 depended on the facts and suggestions retrieved from associative memory. You probably rejected the idea that Republican politics provide protection against kidney cancer. Very likely, you ended up focusing on the fact that the counties with low incidence of cancer are mostly rural. The witty statisticians Howard Wainer and Harris Zwerling, from whom I learned this example, commented, “It is both easy and tempting to infer that their low cancer rates are directly due to the clean living of the rural lifestyle—no air pollution, no water pollution, access to fresh food without additives.” This makes perfect sense.

Now consider the counties in which the incidence of kidney cancer is highest. These ailing counties tend to be mostly rural, sparsely populated, and located in traditionally Republican states in the Midwest, the South, and the West. Tongue-in-cheek, Wainer and Zwerling comment: “It is easy to infer that their high cancer rates might be directly due to the poverty of the rural lifestyle—no access to good medical care, a high-fat diet, and too much alcohol, too much tobacco.” Something is wrong, of course. The rural lifestyle cannot explain both very high and very low incidence of kidney cancer.

The key factor is not that the counties were rural or predominantly Republican. It is that rural counties have small populations. And the main lesson to be learned is not about epidemiology, it is about the difficult relationship between our mind and statistics. System 1 is highly adept in one form of thinking—it automatically and effortlessly identifies causal connections between events, sometimes even when the connection is spurious. When told about the high-incidence counties, you immediately assumed that these counties are different from other counties for a reason, that there must be a cause that explains this difference. As we shall see, however, System 1 is inept when faced with “merely statistical” facts, which change the probability of outcomes but do not cause them to happen.

A random event, by definition, does not lend itself to explanation, but collections of random events do behave in a highly regular fashion. Imagine a large urn filled with marbles. Half the marbles are red, half are white. Next, imagine a very patient person (or a robot) who blindly draws 4 marbles from the urn, records the number of red balls in the sample, throws the balls back into the urn, and then does it all again, many times. If you summarize the results, you will find that the outcome "2 red, 2 white" occurs (almost exactly) 6 times as often as the outcome "4 red" or "4 white." This relationship is a mathematical fact. You can predict the outcome of repeated sampling from an urn just as confidently as you can predict what will happen if you hit an egg with a hammer. You cannot predict every detail of how the shell will shatter, but you can be sure of the general idea. There is a difference: the satisfying sense of causation that you experience when thinking of a hammer hitting an egg is altogether absent when you think about sampling.

A related statistical fact is relevant to the cancer example. From the same urn, two very patient marble counters thatpy dake turns. Jack draws 4 marbles on each trial, Jill draws 7. They both record each time they observe a homogeneous sample—all white or all red. If they go on long enough, Jack will observe such extreme outcomes more often than Jill—by a factor of 8 (the expected percentages are 12.5% and 1.56%). Again, no hammer, no causation, but a mathematical fact: samples of 4 marbles yield extreme results more often than samples of 7 marbles do.

Now imagine the population of the United States as marbles in a giant urn. Some marbles are marked KC, for kidney cancer. You draw samples of marbles and populate each county in turn. Rural samples are smaller than other samples. Just as in the game of Jack and Jill, extreme outcomes (very high and/or very low cancer rates) are most likely to be found in sparsely populated counties. This is all there is to the story.

We started from a fact that calls for a cause: the incidence of kidney cancer varies widely across counties and the differences are systematic. The explanation I offered is statistical: extreme outcomes (both high and low) are more likely to be found in small than in large samples. This explanation is not causal. The small population of a county neither causes nor prevents cancer; it merely allows the incidence of cancer to be much higher (or much lower) than it is in the larger population. The deeper truth is that there is nothing to explain. The incidence of cancer is not truly lower or higher than normal in a county with a small population, it just appears to be so in a particular year because of an accident of sampling. If we repeat the analysis next year, we will observe the same general pattern of extreme results in the small samples, but the counties where cancer was common last year will not necessarily have a high incidence this year. If this is the case, the differences between dense and rural counties do not really count as facts: they are what scientists call artifacts, observations that are produced entirely by some aspect of the method of research—in this case, by differences in sample size.

The story I have told may have surprised you, but it was not a revelation. You have long known that the results of large samples deserve more trust than smaller samples, and even people who are innocent of statistical knowledge have heard about this law of large numbers. But “knowing” is not a yes-no affair and you may find that the following statements apply to you:

- The feature “sparsely populated” did not immediately stand out as relevant when you read the epidemiological story.
- You were at least mildly surprised by the size of the difference between samples of 4 and samples of 7.
- Even now, you must exert some mental effort to see that the following two statements mean exactly the same thing:
- Large samples are more precise than small samples.
- Small samples yield extreme results more often than large samples do.

The first statement has a clear ring of truth, but until the second version makes intuitive sense, you have not truly understood the first.

The bottom line: yes, you did know that the results of large samples are more precise, but you may now realize that you did not know it very well. You are not alone. The first study that Amos and I did together showed that even sophisticated researchers have poor intuitions and a wobbly understanding of sampling effects.

The Law of Small Numbers

My collaboration with Amos in the early 1970s began with a discussion of the claim that people who have had no training in statistics are good “intuitive statisticians.” He told my seminar and me of researchers at the University of Michigan who were generally optimistic about intuitive statistics. I had strong feelings about that claim, which I took personally: I had recently discovered that I was not a good intuitive statistician, and I did not believe that I was worse than others.

For a research psychologist, sampling variation is not a curiosity; it is a nuisance and a costly obstacle, which turns the undertaking of every research project into a gamble. Suppose that you wish to confirm the hypothesis that the vocabulary of the average six-year-old girl is larger than the vocabulary of an average boy of the same age. The hypothesis is true in the population; the average vocabulary of girls is indeed larger. Girls and boys vary a great deal, however, and by the luck of the draw you could select a sample in which the difference is inconclusive, or even one in which boys actually score higher. If you are the researcher, this outcome is costly to you because you have wasted time and effort, and failed to confirm a hypothesis that was in fact true. Using a sufficiently large sample is the only way to reduce the risk. Researchers who pick too small a sample leave themselves at the mercy of sampling luck.

The risk of error can be estimated for any given sample size by a fairly simple procedure. Traditionally, however, psychologists do not use calculations to decide on a sample size. They use their judgment, which is commonly flawed. An article I had read shortly before the debate with Amos demonstrated the mistake that researchers made (they still do) by a dramatic observation. The author pointed out that psychologists commonly chose samples so small that they exposed themselves to a 50% risk of failing to confirm their true hypotheses! No researcher in his right mind would accept such a risk. A plausible explanation was that psychologists' decisions about sample size reflected prevalent intuitive misconceptions of the extent of sampling variation.

The article shocked me, because it explained some troubles I had had in my own research. Like most research psychologists, I had routinely chosen samples that were too small and had often obtained results that made no sense. Now I knew why: the odd results were actually artifacts of my research method. My mistake was particularly embarrassing because I taught statistics and knew how to compute the sample size that would reduce the risk of failure to an acceptable level. But I had never chosen a sample size by computation. Like my colleagues, I had trusted tradition and my intuition in planning my experiments and had never thought seriously about the issue. When Amos visited the seminar, I had already reached the conclusion that my intuitions were deficient, and in the course of the seminar we quickly agreed that the Michigan optimists were wrong.

Amos and I set out to examine whether I was the only fool or a member of a majority of fools, by testing whether researchers selected for mathematical expertise would make similar mistakes. We developed a questionnaire that described realistic research situations, including replications of successful experiments. It asked the researchers to choose sample sizes, to assess the risks of failure to which their decisions exposed them, and to provide advice to hypothetical graduate students planning their research. Amos collected the responses of a group of sophisticated participants (including authors of two statistical textbooks) at a meetatipp>

Amos and I called our first joint article "Belief in the Law of Small Numbers." We explained, tongue-in-cheek, that "intuitions about random sampling appear to satisfy the law of small numbers, which asserts that the law of large numbers applies to small numbers as well." We also included a strongly worded recommendation that researchers regard their "statistical intuitions with proper suspicion and replace impression formation by computation whenever possible."

A Bias of Confidence Over Doubt

In a telephone poll of 300 seniors, 60% support the president.

If you had to summarize the message of this sentence in exactly three words, what would they be? Almost certainly you would choose "elderly support president." These words provide the gist of the story. The omitted details of the poll, that it was done on the phone with a sample of 300, are of no interest in themselves; they provide background information that attracts little attention. Your summary would be the same if the sample size had been different. Of course, a completely absurd number would draw your attention ("a telephone poll of 6 [or 60 million] elderly voters…"). Unless you are a professional, however, you may not react very differently to a sample of 150 and to a sample of 3,000. That is the meaning of the statement that "people are not adequately sensitive to sample size."

The message about the poll contains information of two kinds: the story and the source of the story. Naturally, you focus on the story rather than on the reliability of the results. When the reliability is obviously low, however, the message will be discredited. If you are told that "a partisan group has conducted a flawed and biased poll to show that the elderly support the president…" you will of course reject the findings of the poll, and they will not become part of what you believe. Instead, the partisan poll and its false results will become a new story about political lies. You can choose to disbelieve a message in such clear-cut cases. But do you discriminate sufficiently between "I read in *The New York Times*…" and "I heard at the watercooler…"? Can your System 1 distinguish degrees of belief? The principle of WY SIATI suggests that it cannot.

As I described earlier, System 1 is not prone to doubt. It suppresses ambiguity and spontaneously constructs stories that are as coherent as possible. Unless the message is immediately negated, the associations that it evokes will spread as if the message were true. System 2 is capable of doubt, because it can maintain incompatible possibilities at the same time. However, sustaining doubt is harder work than sliding into certainty. The law of small numbers is a manifestation of a general bias that favors certainty over doubt, which will turn up in many guises in following chapters.

The strong bias toward believing that small samples closely resemble the population from which they are drawn is also part of a larger story: we are prone to exaggerate the consistency and coherence of what we see. The exaggerated faith of researchers in what can be learned from a few observations is closely related to the halo effect thphe , the sense we often get that we know and understand a person about whom we actually know very little. System 1 runs ahead of the facts in constructing a rich image on the basis of scraps of evidence. A machine for jumping to conclusions will act as if it believed in the law of small numbers. More generally, it will produce a representation of reality that makes too much sense.

Cause and Chance

The associative machinery seeks causes. The difficulty we have with statistical regularities is that they call for a different approach. Instead of focusing on how the event at hand came to be, the statistical view relates it to what could have happened instead. Nothing in particular caused it to be what it is—chance selected it from among its alternatives.

Our predilection for causal thinking exposes us to serious mistakes in evaluating the randomness of truly random events. For an example, take the sex of six babies born in sequence at a hospital. The sequence of boys and girls is obviously random; the events are independent of each other, and the number of boys and girls who were born in the hospital in the last few hours has no effect whatsoever on the sex of the next baby. Now consider three possible sequences:

BBBGGG
GGGGGG
BGBBGB

Are the sequences equally likely? The intuitive answer—"of course not!"—is false. Because the events are independent and because the outcomes B and G are (approximately) equally likely, then any possible sequence of six births is as likely as any other. Even now that you know this conclusion is true, it remains counterintuitive, because only the third sequence appears random. As expected, BGBBGB is judged much more likely than the other two sequences. We are pattern seekers, believers in a coherent world, in which regularities (such as a sequence of six girls) appear not by accident but as a result of mechanical causality or of someone's intention. We do not expect to see regularity produced by a random process, and when we detect what appears to be a rule, we quickly reject the idea that the process is truly random. Random processes produce many sequences that convince people that the process is not random after all. You can see why assuming causality could have had evolutionary advantages. It is part of the general vigilance that we have inherited from ancestors. We are automatically on the lookout for the possibility that the environment has changed. Lions may appear on the plain at random times, but it would be safer to notice and respond to an apparent increase in the rate of appearance of prides of lions, even if it is actually due to the fluctuations of a random process.

The widespread misunderstanding of randomness sometimes has significant consequences. In our article on representativeness, Amos and I cited the statistician William Feller, who illustrated the ease with which people see patterns where none exists. During the intensive rocket bombing of London in World War II, it was generally believed that the bombing could not be random because a map of the hits revealed conspicuous gaps. Some suspected that German spies were located in the unharmed areas. A careful statistical analysis revealed that the distribution of hits was typical of a random process—and typical as well in evoking a strong impression that it was not random. "To the untrained eye," Feller remarks, "randomness appears as regularity or tendency to cluster."

I soon had an occasion to apply what I had learned frpeaprainom Feller. The Yom Kippur War broke out in 1973, and my only significant contribution to the war effort was to advise high officers in the Israeli Air Force to stop an investigation. The air war initially went quite badly for Israel, because of the unexpectedly good performance of Egyptian ground-to-air missiles. Losses were high, and they appeared to be unevenly distributed. I was told of two squadrons flying from the same base, one of which had lost four planes while the other had lost none. An inquiry was initiated in the hope of learning what it was that the unfortunate squadron was doing wrong. There was no prior reason to believe that one of the squadrons was more effective than the other, and no operational differences were found, but of course the lives of the pilots differed in many random ways, including, as I recall, how often they went home between missions and something about the conduct of debriefings. My advice was that the command should accept that the different outcomes were due to blind luck, and that the interviewing of the pilots should stop. I reasoned that luck was the most likely answer, that a random search for a nonobvious cause was hopeless, and that in the meantime the pilots in the squadron that had sustained losses did not need the extra burden of being made to feel that they and their dead friends were at fault.

Some years later, Amos and his students Tom Gilovich and Robert Vallone caused a stir with their study of misperceptions of randomness in basketball. The "fact" that players occasionally acquire a hot hand is generally accepted by players, coaches, and fans. The inference is irresistible: a player sinks three or four baskets in a row and you cannot help forming the causal judgment that this player is now hot, with a temporarily increased propensity to score. Players on both teams adapt to this judgment—teammates are more likely to pass to the hot scorer and the defense is more likely to doubleteam. Analysis of thousands of sequences of shots led to a disappointing conclusion: there is no such thing as a hot hand in professional basketball, either in shooting from the field or scoring from the foul line. Of course, some players are more accurate than others, but the sequence of successes and missed shots satisfies all tests of randomness. The hot hand is entirely in the eye of the beholders, who are consistently too quick to perceive order and causality in randomness. The hot hand is a massive and widespread cognitive illusion.

The public reaction to this research is part of the story. The finding was picked up by the press because of its surprising conclusion, and the general response was disbelief. When the celebrated coach of the Boston Celtics, Red Auerbach, heard of Gilovich and his study, he responded, "Who is this guy? So he makes a study. I couldn't care less." The tendency to see patterns in randomness is overwhelming—certainly more impressive than a guy making a study.

The illusion of pattern affects our lives in many ways off the basketball court. How many good years should you wait before concluding that an investment adviser is unusually skilled? How many successful acquisitions should be needed for a board of directors to believe that the CEO has extraordinary flair for such deals? The simple answer to these questions is that if you follow your intuition, you will more often than not err by misclassifying a random event as systematic. We are far too willing to reject the belief that much of what we see in life is random.

I began this chapter with the example of cancer incidence across the United States. The example appears in a book intended for statistics teachers, but I learned about it from an amusing article by the two statisticians I quoted earlier, Howard Wainer and Harris Zwerling. Their essay focused on a large iivepothersnvestment, some $1.7 billion, which the Gates Foundation made to follow up intriguing findings on the characteristics of the most successful schools. Many researchers have sought the secret of successful education by identifying the most successful schools in the hope of discovering what distinguishes them from others. One of the conclusions of this research is that the most successful schools, on average, are small. In a survey of 1,662 schools in Pennsylvania, for instance, 6 of the top 50 were small, which is an overrepresentation by a factor of 4. These data encouraged the Gates Foundation to make a substantial investment in the creation of small schools, sometimes by splitting large schools into smaller units. At least half a dozen other prominent institutions, such as the Annenberg Foundation and the Pew Charitable Trust, joined the effort, as did the U.S. Department of Education's Smaller Learning Communities Program.

This probably makes intuitive sense to you. It is easy to construct a causal story that explains how small schools are able to provide superior education and thus produce high-achieving scholars by giving them more personal attention and encouragement than they could get in larger schools. Unfortunately, the causal analysis is pointless because the facts are wrong. If the statisticians who reported to the Gates Foundation had asked about the characteristics of the worst schools, they would have found that bad schools also tend to be smaller than average. The truth is that small schools are not better on average; they are simply more variable. If anything, say Wainer and Zwerling, large schools tend to produce better results, especially in higher grades where a variety of curricular options is valuable.

Thanks to recent advances in cognitive psychology, we can now see clearly what Amos and I could only glimpse: the law of small numbers is part of two larger stories about the workings of the mind.

- The exaggerated faith in small samples is only one example of a more general illusion—we pay more attention to the content of messages than to information about their reliability, and as a result end up with a view of the world around us that is simpler and more coherent than the data justify. Jumping to conclusions is a safer sport in the world of our imagination than it is in reality.
- Statistics produce many observations that appear to beg for causal explanations but do not lend themselves to such explanations. Many facts of the world are due to chance, including accidents of sampling. Causal explanations of chance events are inevitably wrong.

Speaking of the Law of Small Numbers

“Yes, the studio has had three successful films since the new CEO took over. But it is too early to declare he has a hot hand.”

“I won’t believe that the new trader is a genius before consulting a statistician who could estimate the likelihood of his streak being a chance event.”

“The sample of observations is too small to make any inferences. Let’s not follow the law of small numbers.”

“I plan to keep the results of the experiment secret until we have a sufficiently large sample. Otherwisortpxpere we will face pressure to reach a conclusion prematurely.”