

THE LAW OF SMALL NUMBERS

A study of the incidence of kidney cancer in the 3,141 counties of the United States 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 Zwierling, 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 Zwierling 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 take 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: