tions inform the design and enforcement of policies. If inaccurate tools are used, bad decisions are made—as when a doctor diagnoses a tumor just by measuring body temperature. When the science we're studying is about understanding human trafficking, mass murders, or terrorism, using the best tools and providing the best inputs mean preserving lives.

It's therefore time to retire the idea that understanding crime means understanding the minds and actions of criminals. We must also retire other naïve concepts, such as "organized crime" and the idea that any current nation or government evolves without any criminal influence. These are nicely simplified concepts that work well in theoretical models in the classrooms and journals that manage to evade the complexity and vagueness of society. But if we don't deal with society's true complexity using the diverse tools provided by science, we'll have to deal with that complexity in the streets and the courtrooms—like it or not.

STATISTICAL SIGNIFICANCE

CHARLES SEIFE

Professor of journalism, New York University; former journalist, Science; *author,* Virtual Unreality

It's a boon for the mediocre, the credulous, the dishonest, and the merely incompetent. It turns a meaningless result into something publishable, transforms a waste of time and effort into the raw fuel of scientific careers. It was designed to help researchers distinguish a real effect from a statistical fluke, but it has become a quantitative justification for dressing nonsense up in the mantle of respectability. And it's the single biggest reason that most of the scientific and medical literature isn't worth the paper it's written on.

When used correctly, the concept of statistical significance is a measure to rule out the vagaries of chance—nothing more, nothing less. Say, for example, you're testing the effectiveness of a drug. Even if the compound is completely inert, there's a good chance (roughly 50 percent, in fact) that patients will respond better to your drug than to a placebo. Randomness alone might imbue your drug with seeming efficacy. But the more marked the difference between the drug and the placebo, the less likely it is that randomness alone is responsible. A "statistically significant" result is one that has passed an arbitrary threshold. In most social-science journals and the medical literature, an observation is typically considered statistically significant if there's less than a 5 percent chance that pure randomness can account for the effect you're seeing. In physics, the threshold is usually lower, often 0.3 percent (three sigma) or even

0.00003 percent (five sigma). But the essential dictum is the same: If your result is striking enough to pass that threshold, it's given a weighty label: "statistically significant."

Most of the time, though, the term isn't used correctly. If you look at a typical paper published in the peer-reviewed literature, you'll see that never is just a single observation tested for statistical significance but instead handfuls, or dozens, or even 100 or more. A researcher looking at a painkiller for arthritis sufferers will look at data to answer question after question: Does the drug help a patient's pain? Does it help a patient with knee pain? With back pain? With elbow pain? With severe pain? With moderate pain? With moderate to severe pain? Does it help a patient's range of motion? Quality of life? Each one of these questions is tested for statistical significance, and typically gauged against the industry-standard 5-percent rule. That is, there's a 5-percent chance—1 in 20—that randomness will make a worthless drug seem effective. But test ten questions, and there's a 40-percent chance that randomness will, indeed, deceive you when answering one or more of these questions. And the typical paper asks more than ten questions, often many more. It's possible to correct for this "multiple comparisons" problem mathematically (though it's not the norm to do so). It's also possible to fight this effect by committing to answer just one main question (though in practice such "primary outcomes" are surprisingly malleable). But even these corrections often can't take into account numerous effects that can undermine a researcher's calculations—such as how subtle changes in data classification can affect outcomes. (Is "severe" pain a 7 or above on a 10-point scale, or is it an 8 or above?) Sometimes these issues are overlooked; sometimes they're deliberately ignored or even manipulated.

In the best-case scenario, when statistical significance is calculated correctly, it doesn't tell you much. Sure, chance alone is (relatively) unlikely to be responsible for your observation. But it doesn't reveal anything about whether the protocol was set up correctly, whether a machine's calibration was off, whether a computer code was buggy, whether the experimenter properly blinded the data to prevent bias, whether the scientists truly understood all the possible sources of false signals, whether the glassware was properly sterilized, and so on. When an experiment fails, it's more than likely that the blame rests not on randomness—on statistical flukes—but instead on a good old-fashioned screwup somewhere.

When physicists at CERN claimed to have spotted neutrinos moving faster than light, a six-sigma level of statistical significance (and an exhaustive check for errors) wasn't enough to convince smart physicists that the CERN team had messed up somehow. The result clashed not only with physical law but with observations of neutrinos coming from supernova explosions. Sure enough, a few months later, the flaw (a subtle one) finally emerged, negating the team's conclusion.

Screwups are surprisingly common in science. Consider, for example, the fact that the Food and Drug Administration inspects a few hundred clinical laboratories each year. Roughly 5 percent of inspections come back with findings that the laboratory is engaged in "significant objectionable conditions and practices" so egregious that its data are considered unreliable. Often these practices include outright fraud. Those are just the blindingly obvious problems visible to an inspector; it's hard to imagine that the real number of lab screwups isn't double or triple or quintuple that. What value is there in calling something statistically significant at the 5-percent or 0.3-percent or

even 0.00003-percent level if there's a 10-percent or 25-percent (or more) chance that the data is gravely undermined by a laboratory error? Even the most ironclad findings of statistical validity lose their meaning when dwarfed by the specter of error. Or worse yet, fraud.

Nevertheless, even though statisticians warn against the practice, a one-size-fits-all finding of statistical significance is all too often taken as a shortcut to determine whether or not an observation is credible or a finding is "publishable." As a consequence, the peer-reviewed literature is littered with statistically significant findings that are irreproducible and implausible—absurd observations with effect sizes orders of magnitude beyond what's even marginally believable.

The concept of "statistical significance" has become a quantitative crutch for the essentially qualitative process of whether or not to take a study seriously. Science would be much better off without it.

SCIENTIFIC INFERENCE VIA STATISTICAL RITUALS

GERD GIGERENZER

Psychologist; director, Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, Berlin; author, Risk Savvy

As a young man, Gottfried Wilhelm Leibniz had a beautiful dream: to discover the calculus that could map every single idea in the world into symbols. Such a universal calculus would put an end to all scholarly bickering. Every passionate *Edge* discussion, for one, could be swiftly resolved by dispassionate calculation. Leibniz optimistically estimated that a few skilled people should be able to work the whole thing out in five years. Yet nobody, Leibniz included, has yet found that Holy Grail.

Nonetheless, Leibniz's dream is alive and thriving in the social and neurosciences. Because the object of the dream has not been found, ersatz objects serve in its place. In some fields it's multiple regression, in others Bayesian statistics. But the champ is the null ritual:

1. Set up a null hypothesis of "no mean difference" or "zero correlation." Don't specify the predictions of your own research hypothesis.
2. Use 5 percent as a convention for rejecting the null. If significant, accept your research hypothesis. Report the result as $p < .05$, $p < .01$, or $p. < .001$, whichever comes next to the obtained p-value.
3. Always perform this procedure.