The Mathematics of Making Up Your Mind

With a wee bit of guidance from an obscure eighteenth-century clergyman, statisticians are figuring out how to factor subjective judgments into their objective equations.

BY WILL HIVELY

PHOTOGRAPHS BY IRAIDA ICAZA

TWO OR THREE TIMES a week, while a life hangs in the balance, James Brophy makes a quick decision. Brophy is a cardiologist at Centre Hospitalier de Verdun, in suburban Montreal, which treats about 300 heart attack victims a year. As they arrive, Brophy orders roughly half of them—the ones who made it to the hospital quickly enough—to be injected with one of two clot-busting drugs, streptokinase or tissue plasminogen activator (t-PA). All cardiologists agree that both drugs work well: more than 90 percent of all patients who receive either medication survive. Where they disagree is on the question of which of the drugs they should use. To be sure, thick reports convey the results of clinical trials designed to test the relative merits of the two drugs. But unfortunately the meaning of the data is confusing.

Like every other cardiologist—and, most certainly, like every patient—Brophy would like to know which drug is superior. And to that end, he's waded through a pile of tricky statistics, skirted deep philosophical questions involving how we can know anything at all, and teamed up with Lawrence Joseph, a biostatistician at McGill University. Last year they published a controversial paper advising other doctors how to cut through the statistical fog. To make a rational choice, Brophy and Joseph declared, physicians of the late twentieth century should learn the mental techniques of an obscure eighteenth-century Englishman: the Reverend Thomas Bayes.

Despite his clerical title, the Reverend Thomas Bayes' most enduring work is mathematical, not spiritual. In 1763 he proposed a procedure, known as Bayes' theorem, for evaluating evidence. Early in this century, with the rise of modern statistics—

Bayes' theorem mixed personal beliefs right in with the math. Ignorant ideas or expert opinions, the predilections of monsters or saints—anything at all could go in.

a different set of procedures for evaluating evidence—Bayes' theorem fell out of favor. Recently, however, some researchers have returned to Bayesian ideas.

Mathematicians, by and large, don't find Bayesian procedures very exciting. The people who use them tend to be analysts working on practical problems that require someone to make a risky decision based on imperfect information: evaluating the health risks of radioactive pollutants, for example, even though precise exposure records may be lacking and the effects of low doses are not well understood; or estimating the reliability of backup diesel generators at nuclear power plants, though there have been very few real-life emergencies. One of the Big Three auto companies even paid a statistician good money to design Bayesian software that forecasts warranty claims for new-model cars, although no data yet exist on the long-term performance of those cars.

Bayesian procedures, in theory, are tailor-made for these kinds of "messy" problems, which often involve complex science, uncertain evidence, and quarreling experts—the sort of mess a cardiologist might face when choosing between streptokinase and t-PA. "I've used these drugs," says Brophy, "and participated in clinical trials." But his limited experience didn't count for much, and two large trials, run in 1990 and 1993, one involving some 20,000 patients and the other almost 30,000, proved equivocal. Streptokinase did slightly better in one, t-PA in the other. "Essentially," says Brophy, "they found no big difference between the two drugs."

There *is* one big difference, though. T-PA costs about $1,530 a pop, streptokinase $220. In Canada and Europe, most doctors give streptokinase. In the United States, most doctors give t-PA. "In the States, you might be a lot more worried about whether someone will sue you if you don't use what the literature says is the 'best' drug," speculates Lawrence Joseph.

According to current wisdom, expensive t-PA probably does work better. T-PA, after all, is an enzyme found naturally in blood-vessel linings. Streptokinase, by contrast, is a foreign enzyme derived from streptococcus bacteria, and it can sometimes trigger an immune response. What's more, t-PA acts only at the site of a clot; streptokinase triggers blood-thinning reactions everywhere in the body.

But until a few years ago, the clinical evidence for that supposed superiority was still missing. Then Genentech, t-PA's manufacturer, joined with four other companies in sponsoring a third clinical trial—a *huge* trial this time, with over 40,000 patients—called GUSTO (Global Utilization of Streptokinase and Tissue Plasminogen Activator in Occluded Arteries). When the results were published in 1993, they looked so good for t-PA that the trial's leading researchers declared the drug "clinically superior" to streptokinase on the basis of this trial alone. The earlier trials, they said, had been flawed.

At the time, Brophy, who had gone back to school for his Ph.D. in epidemiology and biostatistics, was studying statistics with Joseph at McGill. When he learned about Bayes' theorem, it changed his way of thinking about such trial results—or rather, it added precision to a way of thinking he'd always used but had previously considered outside the realm of statistics. It mixed personal beliefs right in with the math. Ignorant ideas or expert opinions, the predilections of monsters or saints—anything at all could go in, and Bayes' theorem would turn out a rational conclusion.

According to standard procedures, an analyst should look objectively at the data from any one study. In evaluating a large clinical trial, for instance, he might say that patients taking drug x survived more often than patients taking drug y, so x is better than y. Everyone who looks at the same data should reach the same conclusion. A Bayesian, however, might look at the evidence and think, "Aha! Just as I suspected: it's a toss-up between those drugs." Another Bayesian might decide that y is better than x.

How are such different conclusions possible? Each Bayesian analyst evaluates the same evidence, using Bayes' theorem. Yet each may also bring other information to bear on the problem. According to many Bayesians, statistics should reflect everything we know about a given question—all relevant prior experience. Each analyst must judge, subjectively, which experience is relevant—folklore? similar clinical trials?—and how much that prior evidence should sway belief in the latest results.

Bayes' theorem does not *require* an analyst to weigh evidence subjectively, but it allows him to do so. And that, say critics, shifts the foundation of the analysis from rock-solid mathematics to the quicksand of personal opinion. Detractors call the Bayesian method an exercise in arbitrary thinking—a soft, subjective brand of statistics.

THE REVEREND Thomas Bayes himself is a shady figure. The first time he surfaced as a mathematician, he was already dead. Posthumously, in 1764, the British Royal Society published Bayes' theorem about probabilities. Essentially it was a formula for updating any kind of belief when confronted with new evidence. Bayes described it originally in words that mathematicians and philosophers still struggle to interpret: "The *probability of an event* is the ratio between the value at which an expectation depending on the happening of the event ought to be computed, and the value of

the thing expected upon its happening."

Fortunately, Bayes had an editor. To illustrate how the method worked—how he *thought* it worked—the editor added an appendix containing a charming example: "Let us imagine to ourselves the case of a person just brought forth in this world" and left alone to observe it. "The Sun would, probably, be the first object that would engage his attention; but after losing it the first night he would be entirely ignorant whether he should ever see it again." Our new person, dreading the uncertainty, decides to compute the probability of sunrise.

During his first night, this babe in the woods might decide that the probability of the sun's coming back is not good. That's called a prior probability. Bayes' theorem explains how a babe should update this belief if he runs across new evidence—in this case a sunrise. He starts with his prior probability, adds the new evidence, cranks it all through a computing machine, and out comes an updated "posterior probability"—the new belief.

You can do it yourself with a simple computer: a bowl and some balls. Start out with, say, one white and one black ball in the bowl, representing "sun will return" and "sun will not return." The odds your computer will give for sunrise are the odds of picking a white ball from the bowl.

At the beginning, you have no idea what to expect; your prior belief is completely arbitrary. Some Bayesians recommend starting, always, with even odds—one white ball and one black. Others don't mind more subjective choices. But let's say you start out with a neutral belief in the face of global catastrophe: 50-50 odds that the sun will return. Every time you see a sunrise, you add a white ball to the bowl. After two sightings, the odds look better for a third sunrise: two to one in favor (67 percent probability). After three sightings, the odds are three to one in favor (75 percent probability), and so on. Every day, as the sun keeps returning, you keep raising the probability it will return yet again. After a while, the initial, arbitrary odds hardly matter. The white balls overwhelm the black balls of doubt, just as evidence should always overwhelm superstition.

When should you give up this cumbersome routine and declare sunrise practically dead certain? Whenever it suits you; there's no rule for stopping.

Early in the nineteenth century the great French mathematician Pierre-Simon Laplace translated Bayes' sketchy ideas into usable formulas. With his work, statistical thinking grew out of its mathematical infancy. Some Bayesians, in fact, say their method should probably be called Laplacean. But no matter whose name is attached, Bayes' brand of statistics reigned supreme for a century. Then, near the end of the nineteenth century, the English statistician Sir Ronald Fisher developed simpler and more objective procedures for analyzing data, and by the 1920s almost everyone was using them. With Fisher's methods, a researcher could determine whether the results from any one study were significant. If they were, there would be no need to look at other studies, no need to update an arbitrary prior belief, and no need for the Reverend Thomas Bayes. Everyone could just believe the evidence at hand.

Nowadays when the results of some new study earn the label "statistically significant," we take this as a mathematical seal of approval. It means we can almost certainly believe the new evidence. Fisher's work gave rise to this notion, clarifying the advantage of large trials. If you flip a coin four times and get three heads, is that significant? Would you conclude that the probability of getting heads is 75 percent? Probably not. Those results could well be a fluke—a random long run. If you flip a coin 1,000 times, you would expect most random long runs, such as 3 or even 30 heads in a row, to be balanced by similar long runs of tails. You would expect, overall, results much closer to 50 percent heads and 50 percent tails. The larger your number of coin flips, the more significant your results are likely to be, and the more confident you can be that they are true.

Fisher and several others developed formal tools for calculating significance. One measure of an experiment's significance is called the p-value, another is called the confidence interval, and yet another is popularly known as the margin of error. These are all ways of compar-

When Brophy and Joseph wrote their paper, they wanted to do more than compare two drugs. They wanted to change the way doctors think about clinical trials.

ing the results actually found in the trial with the numbers you would expect from pure chance. The larger a trial, an experiment, or an opinion poll, the smaller your p-value, confidence interval, or margin of error. If you are looking for small differences between drugs, you need a small margin of error, which means a large clinical trial.

Fisher applied his methods to classic probability problems, such as Gregor Mendel's famous experiments with peas. When Mendel wanted to know if wrinkling was an inherited characteristic, he grew smooth and wrinkled peas, cross-bred them, and looked at the second generation. If about three-fourths of the peas in the second generation were wrinkled, that would suggest wrinkling was inherited as a dominant character; if about one-fourth were wrinkled, it was a recessive character. Any other ratios would rule out inheritance, according to Mendel's genetic hypothesis.

These kinds of experiments reduce the statistician's role, essentially, to pea counting. You don't make any judgment about including, say, similar trials with old, wrinkled beans; you just keep counting thousands of peas until you reach some arbitrary level of significance that satisfies all critics. Lawrence Joseph believes that analysts who use these procedures on such cut-and-dried problems "do not in any way, shape, or form need to know anything about anything. They take the data, plug them in, and get answer."

That's fine for peas but not for new cars, which do not roll off the assembly lines with long-term maintenance data ready to plug in. Nor will standard methods work for estimating the reliability of backup diesel generators, which hardly ever get used. And trials with humans, of course, raise questions far more complex and statistically messy than the heritability of smoothness or wrinkles. "Had you asked Fisher to analyze a clinical trial back then," Joseph continues, "he may not have thought his methods were good for that. We'll never know, but the types of problems he was looking at are very different from what Bayesians are looking at today"—problems like choosing between streptokinase and t-PA.

GUSTO SHOWED that when t-PA was administered rapidly and combined with aggressive follow-up therapy, it clearly came out on top: 93.7 percent of patients who received t-PA survived, versus 92.7 percent of those who received streptokinase. A 1 percent difference may seem small, but in cardiology it can mean a lot. In the United States alone, half a million people die of heart attacks every year. Of course, not all those people make it to the emergency room in time for the drugs to work, but if they did, and if 1 percent more of them survived, that would mean 5,000 lives saved. One percent, in fact, happened to be the cutoff point the researchers who conducted the new trial had chosen as proof of t-PA's "clinical superiority." One additional life saved, they said, among every 100 patients injected would justify the higher cost of t-PA. This is, of course, as Joseph grouses, a subjective opinion.

Aside from that quibble, GUSTO met the gold standard for clinical trials: a very large number of patients—41,021 of them—randomly assigned to groups that received one drug or the other. The 1 percent difference looked real. If streptokinase and t-PA were equally effective, you would almost never see a difference in survival rates as large as 1 percent. By the laws of probability, for a trial that large, there was only one chance in 1,000 that t-PA would perform so much better if it was merely as good as streptokinase. So the conclusion seemed unassailable, according to standard, classical statistics: reach for the t-PA, and shell out the extra bucks.

Most practicing physicians would accept GUSTO as decisive. But Brophy could not bring himself to ignore the earlier trials. In his gut, he could not believe t-PA was that much better than streptokinase. "If you're having to put close to 100,000 people into these trials," he says, "you don't need to be a Bayesian analyst to say, jeez, there's probably not a big difference between them, right?"

He and Joseph decided to reanalyze all the data on streptokinase and t-PA. In March 1995 they published their conclusions in the *Journal of the American Medical Association*. Their widely read article, "Placing Trials in Context Using Bayesian Analysis," was more than a comparison of two drugs. It was an all-out pitch for Bayesian methods. Brophy and Joseph wanted to change the way doctors think about clinical trials. Five pages deep in their paper, after some light mathematical skirmishing, they dropped a bombshell.

By any standard significance test, the 1 percent superiority of t-PA seems almost as certain as a law of physics. But according to Brophy and Joseph, the probability of t-PA's being clinically superior is at best 50-50, if you consider GUSTO evidence alone. And if you have any belief at all in earlier results, the odds for clinical superiority drop rapidly to "negligible."

Any doctor who reads the article can start with a subjective prior belief and then use the published data to reach a personal Bayesian conclusion. Brophy and Joseph explain how to do it. One option, for example, is to start with no prior beliefs, like a standard statistician, and accept only the GUSTO results. Surprisingly, this yields no better than a 50 percent probability that t-PA is clinically

superior. That's because t-PA's 1 percent better survival rate has a margin of error. The small size of the margin doesn't matter. What matters is that if you ran the same trial again, with 41,000 new patients, t-PA might surpass 1 percent or fall short. In a truly random clinical trial, the odds are 50-50 that a new trial would go either way.

Since the GUSTO result of 1 percent is also the cutoff point for t-PA's clinical superiority, the odds are only about 50-50 that t-PA is actually clinically superior, on the basis of GUSTO alone. A non-Bayesian analyst could have figured this out. If the GUSTO researchers had picked any other value for clinical superiority, the odds would have come out differently. As it was, the researchers chose a small value for clinical superiority and yet barely squeaked in with their results. So the strongest degree of belief anyone can reasonably have in t-PA's minimal superiority is merely, shall we say, halfhearted.

This discouraging conclusion does not contradict the impressively high significance of the GUSTO study. Significance is one thing and clinical superiority is another, although it's easy to confuse them—as Brophy suspected some doctors would do when reading the GUSTO study. Given the GUSTO results, say Brophy and Lawrence, the probability is that if you ran 1,000 trials, t-PA might do better 999 times. But how often would it do *1 percent* better? The answer, at best, is about half those times. You can be 99.9 percent certain that t-PA is better than streptokinase and, at the same time, only 50 percent certain it's clinically superior.

That's if you start with no prior belief, like a babe in the woods. If you assign more credibility—any whatsoever—to earlier trials, the clinical superiority of t-PA looks less likely. Brophy and Joseph illustrate three options: 10 percent, 50 percent, and 100 percent belief in the results from two earlier trials. You can choose any degree of belief as your starting point—it's your own subjective judgment.

If you choose 10 percent, it means you are giving the earlier evidence only one-tenth the statistical weight of the GUSTO evidence. If you accept earlier results at the maximum value, 100 percent, you are skeptical that the differences between tri-

Starting out with subjective beliefs is the whole point of Bayesian statistics; it liberates analysts from the "statistical slavery" of bean counting.

als—the rapid administration of t-PA, the follow-up therapy, and so on—mean very much. You are willing to take the results from all three trials at face value and lump them together. That option yields the lowest probability, nearly zero, that t-PA is clinically superior.

But why stop at an impartial, equal belief in all three trials? Brophy and Joseph drop a hint that the GUSTO data might count for less. Physicians participating in the study knew which drug they were giving—it was not a blind trial—and patients who got t-PA were "apparently" 1 percent more likely to have a coronary bypass operation as well.

THIS KIND of give-and-take is typical of complex scientific problems. Contradictory conclusions are also typical; you can see them every day in newspapers. On January 4, for example, the Associated Press wire service reported this new evidence on global warming: "The world's average surface temperature for 1995 was 58.72 degrees Fahrenheit, or 0.7 degree higher than the average for 1961–1990, said Phil Jones of the Climatic Research Unit at the University of East Anglia in England." It was the highest average surface temperature ever recorded for a single year.

Three experts commented on the new statistic. " 'This is the culmination of a whole series of findings which demonstrate that the world is warming,' said Michael Oppenheimer, an atmospheric scientist at the Environmental Defense Fund. 'The question is no longer whether the climate is changing, the question is now what are we going to do about it.' " Oppenheimer's prior belief in other evidence clearly influenced him to accept this new information about surface temperature at face value.

Kevin Trenberth of the National Center for Atmospheric Research in Boulder, Colorado, "cautioned that the British study may exaggerate the amount of overall warming." It seemed to him perhaps a fluke. This could be an argument for counting more peas.

The third expert, climatologist Patrick Michaels of the University of Virginia, referred to specific prior evidence. " 'There is now a statistically significant difference,' he said, 'between the temperatures measured in that [British] land-based record and temperatures measured by satellites.' " He gave the satellite evidence more weight; satellites have better coverage. " 'The net temperature trend in the satellite record, which just finished its 17th year, is actually slightly negative,' he said. 'I think in the long run you're just going to see increasing confirmation of the hypothesis that the warming would not be nearly as large as it was forecast to be.' "

To Joseph, this all represents ad hoc interpretation that would be vastly improved by Bayesian analysis. Of course, no one expects a researcher to reel off impeccable Bayesian logic while talking to a reporter. But Joseph sees a deeper problem, which demonstrates the glaring deficiency of standard procedures. First some "objective" data arrive on the scene, almost completely out of context. Then some "interpretation" smacks the new evidence around. This happens even in scientific journals. Researchers pull prior evidence from all over the map, "but it's never done at a formal level," Brophy complains. "The difference in a Bayesian analysis is that it forces you to consider the prior information formally, and formally put it into your analysis." That way, "at least you could check the subjectivity." And subjectivity always exists, or scientists would never disagree. "It is the degree that subjectivity is *apparent*," Joseph says, "that makes for good science."

Most Bayesians would say that your prior "degrees of belief" should not be pulled from thin air. Some say they should not be pulled from anywhere; since they believe Bayesian statistics should be no more subjective than the classical kind, they favor using a standard value, such as giving each set of data equal weight. Others, like George Washington University statistician Nozer Singpurwalla, one of the more passionate Bayesian revivalists, say that starting out with subjective degrees of belief is the whole point of Bayesian statistics; it liberates analysts from the "statistical slavery" of bean counting.

As for Brophy, he is still at Centre Hospitalier de Verdun, going about his grueling business. In his spare time he is also still working on his Ph.D. Along the way, as part of a project for only one class, he may have changed the way clinicians will interpret data from clinical trials—or at least one of the largest trials in medical history. He personally rates the chances of t-PA's being clinically superior to streptokinase as "no better than 5 or 10 percent." At that rate, t-PA might save only one more life among every 250 heart attack victims. Would that person's life be worth the extra $327,500, say, that it would cost to give all those patients the more expensive medicine?

"As a physician," Brophy says, "your primary responsibility is toward your patient. You also have to give a little thought to the next patients coming in. Maybe your hospital is going broke and can't treat them down the line. These are hard questions to face. People would rather not."

And that, in the end, may be the biggest problem the Bayesians face. Their procedure *forces* people to examine their beliefs, as Joseph says—and possibly change them. People would rather not.