Bayes Offers a `New' Way to Make Sense of Numbers.(Bayesian statistics)

by DAVID MALAKOFF

A 236-year-old approach to statistics is making a comeback, as its ability to factor in hunches as well as hard data finds applications from pharmaceuticals to fisheries

After 15 years, environmental researcher Kenneth Reckhow can still feel the sting of rejection. As a young scientist appearing before an Environmental Protection Agency review panel, Reckhow was eager to discuss his idea for using an unorthodox statistical approach in a water-quality study. But before he could say a word, an influential member of the panel unleashed a rhetorical attack that stopped him cold. "As far as he was concerned, I was a Bayesian, and Bayesian statistics were worthless," recalls Reckhow, now at Duke University in Durham, North Carolina. "The idea was dead before I even got to speak."

Reckhow is no longer an academic outcast. And the statistical approach he favors, named after an 18th century Presbyterian minister, Thomas Bayes, now receives a much warmer reception from the scientific establishment. Indeed, Bayesian statistics, which allows researchers to use everything from hunches to hard data to compute the probability that a hypothesis is correct, is experiencing a renaissance in fields of science ranging from astrophysics to genomics and in real-world applications such as testing new drugs and setting catch limits for fish. The long-dead minister is also weighing in on lawsuits and public policy decisions (see p. 1462), and is even making an appearance in consumer products. It is his ghost, for instance, that animates the perky paperclip that pops up on the screens of computers running Microsoft Office software, making Bayesian guesses about what advice the user might need. "We're in the midst of a Bayesian boom," says statistician John Geweke of the University of Iowa, Iowa City.

Advances in computers and the limitations of traditional statistical methods are part of the reason for the new popularity of this old approach. But researchers say the Bayesian approach is also appealing because it allows them to factor expertise and prior knowledge into their computations--something that traditional methods frown upon. In addition, advocates say it produces answers that are easier to understand and forces users to be explicit about biases obscured by reigning "frequentist" approaches.

To be sure, Bayesian proponents say the approach is no panacea--and the technique has detractors. Some researchers fear that because Bayesian analysis can take into account prior opinion, it could spawn less objective evaluations of experimental results. "The problem is that prior beliefs can be just plain wrong" or difficult to quantify properly, says statistician Lloyd Fisher of the University of Washington, Seattle. Physicians enthusiastic about a particular treatment, for instance, could subtly sway trial results in their favor. Even some advocates worry that increased use may lead to increased abuse. "There is a lot of garbage masquerading under the Bayesian banner," warns statistician Don Berry of the University of Texas M. D. Anderson Cancer Center in Houston, a leading advocate of Bayesian approaches.

Still, the renewed interest in Bayesian methods represents a major upturn in a 2-century roller-coaster ride for this approach to data analysis. Two years after Bayes's death in 1761, a friend, Richard Price, arranged for the British Royal Society to publish his notes on "a problem in the doctrine of chances." The 48-page essay tackled a question that was as much philosophy as mathematics: How should a person update an existing belief when presented with new evidence, such as the results from an experiment? Bayes's answer was a theorem that quantified how the new evidence changed the probability that the existing belief was correct (see p. 1461). The theorem "is mathematics on top of common sense," says statistician Kathryn Blackmond Laskey of George Mason University in Fairfax, Virginia.

Using the theorem for any but the simplest problems, however, was beyond the skills of most mathematicians of that era. As a result, few scientists were aware of Bayes's ideas until the 1790s, when the French mathematician Pierre-Simon de Laplace showed researchers how they could more easily apply them. Laplace's work came to dominate applied statistics over the next century. Still, some scientists became increasingly uneasy about one characteristic of the Bayesian-Laplacian mathematical framework: Two people analyzing the same evidence can arrive at dramatically different answers if they start with different beliefs and experiences.

Researchers eager to avoid that problem got their wish in the 1930s, after statisticians Ronald A. Fisher and Egon Pearson of the United Kingdom and Jerzy Neyman of Poland offered new methods of evaluating data and comparing competing hypotheses. The intertwined methods became known as frequentist statistics because they indicate how frequently a researcher could expect to obtain a given result if an experiment were repeated and analyzed the same way many times. Frequentist methods were viewed as more objective because different researchers could apply them to the same data and usually emerge with similar answers, no matter what beliefs they started with.

Frequentist techniques had another advantage: They proved relatively easy to apply to real-world problems--unlike Bayesian methods. "It might take just half an hour to write down the equation" needed to answer a problem with Bayesian statistics, explains Brian Junker, a statistician at Carnegie Mellon University in Pittsburgh, Pennsylvania--"but forever to do the computation." As a result, adds Greg Wilson, a doctoral student in rhetoric at New Mexico State University in Las Cruces, "frequentists used to say to Bayesians, `You're wrong--but even if you weren't wrong, you still can't do the computation.'"

That argument, however, began to dissolve earlier this decade with the growing power of desktop computers and the development of new algorithms, the mathematical recipes that guide users through problems. Bayesian statisticians, including Alan Gelfand of the University of Connecticut, Storrs, Adrian Smith of Imperial College, London, and Luke Tierney of the University of Minnesota, Minneapolis, helped popularize the use of simulation techniques now known as Markov Chain Monte Carlo--or "MCMC" to insiders. The new tools made the Bayesian approach accessible to a wide range of users, who say it has significant advantages. One is that it allows researchers to plug in prior knowledge, whereas frequentist approaches require users to blind themselves to existing information because it might bias the results.

Such prior information can be very helpful to researchers trying to discern patterns in massive data sets or in problems where many variables may be influencing an observed result. Larry Bretthorst of Washington University in St. Louis, Missouri, for instance, developed Bayesian software that improved the resolution of nuclear magnetic resonance (NMR) spectrum data--used by chemists to figure out the structure of molecules--by several orders of magnitude. It uses prior knowledge about existing NMR spectra to clarify confusing data, yielding resolution improvements that were "so startling that other researchers had a hard time believing he hadn't made a mistake," says Kevin Van Horn, an independent computer scientist in American Fork, Utah.

Genomics researchers have also become converts. "You just say `Bayesian,' and people think you are some kind of genius," says statistician Gary Churchill of The Jackson Laboratory in Bar Harbor, Maine, who is working on ways to analyze the flood of data produced by DNA sequencing and gene expression research. Some researchers, for instance, are using what they already know about a DNA sequence to identify other sequences that have a high probability of coding for proteins that have similar functions or structures, notes Jun Liu, a statistician at Stanford University in Palo Alto, California. "No easy frequentist method can achieve this," he says. Similarly, "Bayesian has become the method of choice" in many astrophysics studies, says astrophysicist Tom Loredo of Cornell University in Ithaca, New York. The approach has allowed users to discern weak stellar signal patterns amid cosmic background noise and take a crack at estimating the locations and strengths of mysterious gamma ray bursts.

Lifesaving statistics?

In other fields, such as drag and medical device trials, Bayesian methods could have practical advantages, say advocates. Indeed, at a 2-day conference last year, the Food and Drug Administration (FDA) office that approves new devices strongly urged manufacturers to adopt Bayesian approaches, arguing that they can speed decisions and reduce costs by making trials smaller and faster.

Telba Irony, one of two Bayesian statisticians recently hired by the division, says the savings flow from two advantages of the Bayesian approach--the ability to use findings from prior trials and flexibility in reviewing results while the trial is still running. Whereas frequentist methods require trials to reach a prespecified sample size before stopping, Bayesian techniques allow statisticians to pause and review a trial to determine--based on prior experience--the probability that adding more patients will appreciably change the outcome. "You should be able to stop some trials early," she says. So far, just a handful of the 27,000 device firms regulated by FDA have taken advantage of the approach. But FDA biostatistician Larry Kessler hopes that up to 5% of device trials will be at least partly Bayesian within a few years. "We're not going to change the statistical paradigm overnight," he says. "There is still a healthy degree of skepticism out there."

Such skepticism has also limited the use of Bayesian approaches in advanced drug trials, a potentially much bigger arena. But a team led by M. D. Anderson's Berry and researchers at Pfizer Inc.'s central research center in Sandwich, England, is about to challenge that taboo. Next June, using a heavily Bayesian study design, the company plans to begin human trials aimed at finding the safest effective dose of an experimental stroke drug designed to limit damage to the brain. The trial--called a phase II dose ranging trial--will help the company decide whether to move the drug into final testing trials. "There are huge economic consequences on the line" says Pfizer statistician Andy Grieve.

Bayes Offers a `New' Way to Make Sense of Numbers.(Bayesian statistics)

The team believes that Bayesian methods will allow the company to reach conclusions using 30% fewer patients than in traditional designs. Just as important, however, Berry and his colleagues believe that Bayesian flexibility could reduce an ethical problem that they say plagues frequentist-oriented trials: the need to deny beneficial treatments to some patients in the name of statistical rigor: Traditional dose ranging trials, for instance, randomly assign patients to one of several drug doses until enough patients have been treated to produce statistically significant results. Under classic frequentist designs, researchers aren't supposed to look at the data until the study is done, for fear of biasing the analysis. Although many statisticians bend the rules to reduce ethical concerns, many patients in traditional trials may still receive a less beneficial treatment even if, unbeknownst to the investigators, the evidence is mounting that another dose is better. "The frequentist approach inadvertently gives rise to an attitude that we have to sacrifice patients to learn," says Berry, who believes the practice has unnecessarily cost lives in some trials.

The Pfizer study, in contrast, should allow researchers to analyze the accumulating data and more quickly eliminate ineffective or potentially harmful doses. Skeptics worry that the approach could allow the drug-maker's enthusiasm for its product to color the results. But Grieve says the team will also perform more traditional frequentist analyses to satisfy FDA officials and company executives that Bayesian methods are "robust and ready."

Another advantage of the Bayesian approach, say statisticians from both camps, is that it produces answers that are easier to understand than those produced by frequentist computations. These methods generate a measure of uncertainty called the "P value," which researchers find handy. In simple terms, a P value is supposed to tell a researcher whether experimental results are statistically "significant" or the product of chance. For decades, many journals would only publish results with a P value of less than 0.05. In a trial comparing a new drug to no treatment, for instance, a result with a $P = 0.05$ means that the odds that the "null hypothesis"--no treatment--would produce the observed effect are just 1 in 20 if the experiment is repeated many times, suggesting that the alternative hypothesis--the new drug--is creating the effect.

But there are at least two problems with P values, physician and biostatistician Steven Goodman of The Johns Hopkins University in Baltimore, Maryland, noted earlier this year in a plea for greater use of Bayesian methods published in the Annals of Internal Medicine (15 June, p. 995). One is that many people, even those with some statistical training, incorrectly interpret $P = 0.05$ to mean that there is a 95% chance that the null hypothesis is wrong and the alternative hypothesis correct. This misinterpretation exacerbates a second problem: that P values tend to overstate the strength of the evidence for a difference between two hypotheses. Indeed, studies with small P values, implying a highly significant finding, have sometimes paled in a Bayesian reanalysis. For instance, in a widely cited 1995 Bayesian restudy of findings that one heart attack drug worked better than another, Canadian researchers Lawrence Joseph and James Brophy concluded that the use of P values overstated the superiority of one of the drugs. Similarly, Duke's Reckhow found that a statistically significant trend in acid rain pollutants detected in some lakes by frequentist analyses disappeared upon a Bayesian reexamination.

Goodman says that even he is sometimes at a loss to explain the proper interpretation of P values to his students, but that he has no problem getting them to understand Bayesian probabilities. "Bayesian computations give you a straightforward answer you can understand and use," he says. "It says there is an X% probability that your hypothesis is true--not that there is some convoluted chance that if you assume the null hypothesis is true, you'll get a similar or more extreme result if you repeated your experiment thousands of times. How does one interpret that?"

Uncertain decisions

Statisticians say that knowing just how seriously to take the evidence can be particularly important in politics and business, where decisions have to be made in spite of uncertainty. Fisheries managers, for instance, are on the spot to decide how many fish people should be allowed to catch from a population of indeterminate size. Last year, in a bid to improve quota-setting, a National Academy of Sciences panel recommended that fisheries scientists "aggressively" pursue the use of Bayesian methods to predict fish populations and "evaluate alternative management policies." The idea, says panel member Ray Hilborn of the University of Washington, Seattle, is that fisheries researchers should spell out the amount of uncertainty that goes into their predictions. Policy-makers might be more cautious in setting catch quotas, he and others reason, if they knew there was a significant risk that their actions might destroy a stock.

Although the new approach has yet to make "a real difference" in fisheries management, says Hilborn, Bayesian techniques are already influencing policy and business decisions in other fields. The model that U.S. officials use to estimate the population of endangered bowhead whales--and the number that Alaskan natives are allowed to kill--is Bayesian. A Bayesian model is also at the heart of the controversy over whether to use statistical adjustments to avoid under-counts of urban minorities in the next U.S. census; it uses assumptions drawn from intensively surveyed neighborhoods to estimate how many people census-takers may have missed in other areas. Oil, power, and banking companies also call on Bayesian statisticians regularly to help them predict where they should drill, expect electricity demand, or move their investments, and salvage companies and the Coast Guard use Bayesian search models to decide where to look for shipwrecks and lost mariners.

The most ubiquitous Bayesian application, however, may be Microsoft's animated paperclip, which offers help to users of its Office software. The application, says computer scientist Eric Horvitz of Microsoft Research in Redmond, Washington, grew out of software he and colleagues designed that tries to predict what users will ask next by keeping track of prior questions. He can even explain why the paperclip pops up when it isn't wanted. Company officials "didn't go all the way with Bayes--they could have avoided that problem if they had," he claims.

Bayesian barriers

Corporate cautiousness isn't the only factor limiting the spread of Bayesian statistics, however. Another is its absence from the undergraduate curriculum, advocates say. Two years ago, the American Statistician, a journal of the American Statistical Association, published a heated debate about whether it should be included--a debate that continues in e-mail groups run by statisticians. Opponents argue that teaching Bayes would not prepare students for the kinds of statistical applications they are likely to encounter in their professional lives, says David Moore of Purdue University in West Lafayette, Indiana. Supporters respond that the world is changing, and that Bayesian techniques will soon become essential tools.

Ecologists, for instance, are increasingly being drawn into policy debates that push them to state the probable outcomes of different environmental policies, notes Aaron Ellison of Mount Holyoke College in South Hadley, Massachusetts. To prepare his students, he teaches Bayesian methods in his introductory courses and last year edited a special issue of the journal Ecological Applications encouraging colleagues to verse themselves in Bayesian methods.

Also limiting the use of Bayesian tools is the absence of "plug and play" software packages of the kind that have made frequentist approaches so easy to apply. Although several companies are designing products to fill the niche, not all statisticians believe they are a good idea. Bayesian methods are complicated enough, says statistician Brad Carlin of the University of Minnesota, Minneapolis, that giving researchers user-friendly software could be "like handing a loaded gun to a toddler; if the data is crap, you won't get anything out of it regardless of your philosophical bent."

The growing demand for Bayesian aids, however, reflects a profound change in the acceptance of Bayesian methods--and an end to the old debates, says Rob Kass, head of Carnegie Mellon's statistics department. "In my view, Bayesian and frequentist methods will live side by side for the foreseeable future."

For some Bayesians, that is a thrilling notion. George Mason's Laskey and others are even planning a London celebration in 2001 to kick off a "century of Bayes" (decision.iet.com/CoBC/CoBC_ie.htm). Meanwhile, others are savoring the Reverend's return to respectability. "Twenty-five years ago, we used to sit around and wonder, `When will our time come?'" says mathematician Persi Diaconis, a Bayesian at Stanford. "Now we can say: `Our time is now.'"

[ILLUSTRATION OMITTED]

RELATED ARTICLE: A Brief Guide to Bayes Theorem

In making sense of new data, traditional statistical methods ignore the past. Bayesian techniques, in contrast, allow you to start with what you already believe and then see how new information changes your confidence in that belief. Putting Bayes theorem to work requires heavy calculus--now aided by computer models--but the fundamental concept is straightforward.

Thomas Bayes's original 1763 paper, which focused on predicting the behavior of billiard balls, included his now-famous theorem (see center box) and an illustrative example. Imagine, it says, that a newborn arrives on Earth and watches the sun set. The precocious child scientist wants to know the probability that the sun will rise again. Toiling in the darkness with Bayes's theorem, the clueless researcher first must make a guess about what will happen and assign it a probability. So she hedges her bets--essentially admitting that she doesn't know the answer--and decides that there's an even chance, I in 2, that the sun will return. This is called the "prior probability."

To analyze her data, the skywatcher constructs a simple computer using a bag and a collection of marbles, either white or black. She drops one white marble into the bag to represent the likelihood that "the sun will return," and one black marble for the opposite outcome. After each rosy dawn, she adds another white marble to the bag. The changing contents of the bag demonstrate the improving odds in favor of another sunrise, from 2:1, to 20:1, to 100:1. As the experiment continues, the black marble becomes little more than a statistical speck amid a sea of white.

Looking into her bag, the researcher concludes that the "posterior probability" of the sun rising is very high--nearly 100%. Indeed, the new evidence is so strong that it leads her to select one of her equally weighted prior guesses--the sun will rise--over the other.

In theory, Bayesians say enough evidence will lead people who start with dramatically different priors to essentially the same posterior answer. When data are scarce, however, the choice of a prior probability can heavily influence the posterior. That process has led to charges that the method is too subjective.

But even if the final answers differ, Bayesians say, the method can still probe the strength of the evidence itself, by computing how far it alters each prior belief. Modern Bayesians have also tried to reduce the influence of priors by using "flat priors" that avoid playing favorites by giving every possible hypothesis equal probability. Another technique for guarding against bias is "sensitivity testing," in which a wide range of priors is tested to see which one holds up best.

The theorem

$P(H|D) = P(D|H) \times P(H)/P(D)$ says that the probability of the hypothesis, given the data, is equal to the probability of the data, given that the hypothesis is correct, multiplied by the probability of the hypothesis before obtaining the data divided by the averaged probability of the data.

--D.M.

RELATED ARTICLE: The Reverend Bayes Goes to Court

A statistical approach developed by an 18th century clergyman helped the New Jersey Public Defender's Office deal with a thoroughly modern challenge: proving allegations that state troopers were engaged in "racial profiling"--discriminating against African-American drivers by singling them out for traffic stops. The 1994 case, settled this year, was just one of several recent high-profile court appearances for Bayesian statistics, which allows data to be analyzed in the light of prior knowledge--and gives answers that judges, jurors, and other nonstatisticians can more easily interpret.

In New Jersey, the approach enabled the legal team to overcome a huge gap in police records to make their case. The problem was that in 70% of 892 analyzed stops along the southern end of the New Jersey Turnpike, police records did not identify the race of the driver. The lawyers worried that the missing data would doom their case, which sought to suppress evidence against 17 African-American suspects, many of whom were charged with transporting drugs.

But statisticians Joseph Kadane of Carnegie Mellon University in Pittsburgh and Norma Terrin of the New England Medical Center in Boston used a Bayesian model and prior information to make sense of the missing data.(*) They knew, for instance, that based on he 30% of stops where the driver's race was known, blacks appeared to be 4.86 times more likely to be pulled over than drivers of other races. The question was whether troopers were also more likely, to record the race of black drivers. If so, the 70% of stopped drivers of unknown race might have been overwhelmingly white, and the total numbers might not support the profiling charge.

Using prior knowledge, Kadane and Terrin were able to test this possibility. Earlier studies on the turnpike showed that about 15% of speeding drivers were black. For their analysis, the statisticians assumed that the police were as much as three times more likely to record the race of a black driver. Even under such extreme circumstances, they concluded that black drivers were still far more likely to be stopped than white drivers. Last summer, after losing the case and then withdrawing an appeal, state officials began dismissing the 17 cases.

Although a clever statistician might have arrived at a similarly convincing answer with conventional techniques, says Kadane, it would have taken more work. Also, the logic of a Bayesian answer--which gives the probability that the answer lies within a certain range of possible answers--is "more accessible to the court" than hard-to-follow frequentist answers, Kadane and Terrin say.

Bayesian clarity was also critical to undermining a doping accusation against U.S. long-distance runner and former record holder Nary Decker Slaney. Slaney was barred from the sport after failing a drug test for performance-enhancing steroids at the 1996 Olympic games. But the test, which measures the ratio of the steroids testosterone to epitestosterone (T/E) in urine, is statistically impossible to interpret as evidence of guilt or innocence, says statistician Don Berry of the University of Texas M. D. Anderson Cancer Center in Houston, who helped Slaney's defense team win a reversal from U.S. track officials.

Berry used a Bayesian approach to show that prior knowledge is essential to conclude anything from the test. For one, he says, testers need to know "the probability that

Bayes Offers a `New' Way to Make Sense of Numbers.(Bayesian statistics)

you'd get a T/E greater than 6 [the prohibited level] if she used a banned substance, and the same probability if she didn't use it." Doctors who rely on tests to diagnose everything from AIDS to paternity face similar problems, he notes, as do prosecutors trying to pin convictions on other kinds of drug tests. In each case, other statisticians say, the Bayesian framework exposes the range of prior information needed to interpret test results properly.

Whatever the final outcome of Slaney's case--an international federation has so far upheld the ban, and her lawyers are still suing to clear her name--Berry's analysis and other criticism have severely damaged the credibility of the T/E test. It's "bad science, bad policy," physician Gary Wadler of the New York University School of Medicine last month told a congressional committee that is pushing for changes in drug testing among athletes. International Olympic officials still disagree, but the Reverend Bayes may force them to make amends.

(*) "Missing Data in the Forensic Context," Journal of the Royal Statistical Society 160(2), 351 0997).

--D.M.

RELATED ARTICLE: An Improbable Statistician

In 1969, statisticians around the world sent funds to restore the deteriorating London tomb of Thomas Bayes, "in recognition of [his] important work in probability." But beyond his theorem, which opens a place for preconceptions and prior knowledge in statistical analysis, historians know very little about this Presbyterian minister and how he became interested in his "problem in the doctrine of chances."

Born into an affluent family in 1701 or 1702, Bayes followed his father into the pulpit and became a minister in the Nonconformist sect, which had broken from the Church of England. Barred from Oxford or Cambridge because of his religious affiliation, he studied logic and theology for a few years at the University of Edinburgh in Scotland, says David Bellhouse, a Canadian statistician and Bayes historian at the University of Western Ontario. He retired from his ministerial duties in 1752 and died in 1761.

Bayes's notebooks--recently found at an insurance company he helped fund--suggest that he "was obviously well read and kept up on the mathematical literature," Bellhouse says. Indeed, a 1736 book defending Isaac Newton's ideas may have been what won him election to the Royal Society in 1742.

Bayes's two scholarly papers, however, weren't published until after his death and attracted little interest at the time. Indeed, one historian of the

Royal Society later dismissed the probability paper, saying "the solution is much too long and intricate to be of much practical utility." His public profile was so low, in fact, that a widely circulated portrait of the great man is in all likelihood not Bayes at all. "The clothes and hairstyle are all wrong," Bellhouse says.

But more than 2 centuries later, his ideas live on. And what would Bayes have thought about the extended debate over his methodology? "We would like to think he is a subjectivist fellow-traveler," wrote statisticians Jose Bernardo and Adrian Smith in 1994. Luckily, they note, "he is in no position to complain at the liberties we take with his name."

--D.M.