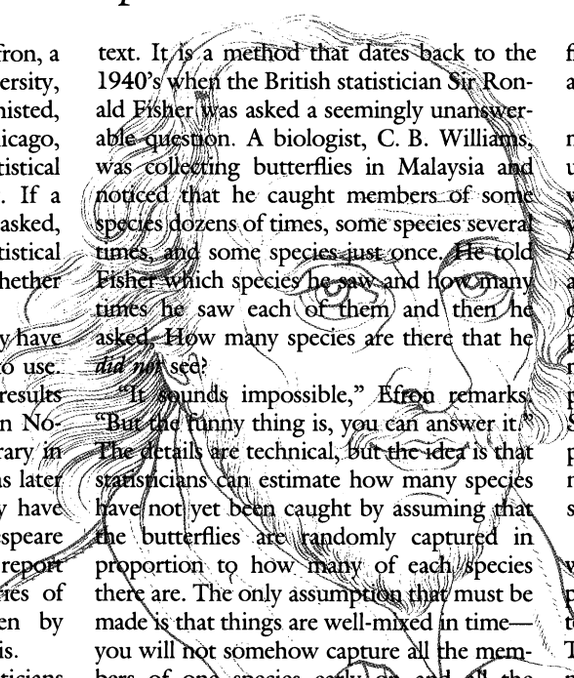
Shakespeare's New Poem: An Ode to Statistics

Two statisticians are using a powerful method to determine whether Shakespeare could have written the newly discovered poem that has been attributed to him



About 10 years ago, Bradley Efron, a statistician at Stanford University, and his student Ronald Thisted, who is now at the University of Chicago, decided, just for fun, to do a statistical analysis of Shakespeare's vocabulary. If a new work were to be found, they asked, how could you determine, by a statistical analysis of the language used, whether Shakespeare could have written it?

Now, to their utter amazement, they have an opportunity to put their analysis to use. Thisted and Efron are applying their results to the poem found by Gary Taylor in November of 1985 in the Bodelain Library in Oxford, England. (A second copy was later found in Yale's library.) So far, they have found no reason to believe that Shakespeare *could not* have written it. And they report that poems by several contemporaries of Shakespeare are clearly not written by Shakespeare, according to this analysis.

This is not the first time that statisticians have used their craft to address literary questions. And, frequently, a statistical analysis is so convincing that it tips the scale—a long-standing literary dispute is settled. For example, Frederick Mosteller and David Wallace of Harvard University used statistics to determine that James Madison and not Alexander Hamiliton is the author of *The Federalist Papers*. In the 1950's, the British statistician David Cox and literary scholar L. Brandwood used statistics to settle a 1000-year-old debate over the order in which Plato wrote his books.

Shall I die? Shall I fly
Lovers' baits and deceits,
sorrow breeding?
Shall I tend? Shall I send?
Shall I sue, and not rue
my proceeding?
In all duty her beauty
Binds me her servant for ever,
If she scorn, I mourn,
I retire to despair, joying never.

But although the idea of using statistics to answer literary questions is not new, the particular method that Thisted and Efron employed has never been used in this context. It is a method that dates back to the 1940's when the British statistician Sir Ronald Fisher was asked a seemingly unanswerable question. A biologist, C. B. Williams, was collecting butterflies in Malaysia and noticed that he caught members of some species dozens of times, some species several times, and some species just once. He told Fisher which species he saw and how many times he saw each of them and then he asked, How many species are there that he *did not* see?

"It sounds impossible," Efron remarks. "But the funny thing is, you can answer it." The details are technical, but the idea is that statisticians can estimate how many species have not yet been caught by assuming that the butterflies are randomly captured in proportion to how many of each species there are. The only assumption that must be made is that things are well-mixed in time—you will not somehow capture all the members of one species early on and all the members of another species later. If a species has not yet been caught it is a matter of bad luck and not systematic evasion. There is no cache of butterfly species hiding somewhere.

Efron and Thisted decided to apply this sort of analysis to Shakespeare's poetry when they heard a lecture by statistician Joseph Gani of the University of California at Santa Barbara. Gani's goal was to analyze the structure of Shakespeare's writings, but in the course of his talk he pointed out that the necessary data were there for another sort of analysis. Marvin Spevack of Westfälische Wilhelms-Universität in Münster had put all of Shakespeare's works on a computer and had used the computer to count all of the words that Shakespeare used, and how many times he used each word in his published works.

It was a situation that resembled the butterfly collections. Here is a collection of 884,647 total words that Shakespeare used in all his known works, consisting of a total vocabulary of 31,534 different words. Of the words in Shakespeare's vocabulary, 14,376 appeared just once, 4,343 appeared just twice, 2,292 appeared just three times, 1,463 appeared four times, 1,043 appeared five times, 837 appeared six times, 638 appeared seven times, and so on.

Now, asked Efron and Thisted, how many words did Shakespeare know but *not* use? If there were a newly discovered poem written by Shakespeare, How many words would be in it that he had never used before? And how many of the words in it would be among those he had previously used only once before? How many would he have previously used only twice before? How many three times? They predicted, for example, that the discovery of a new volume of Shakespeare's works equal in length to all his previous works would contan 11,430 ± 178 new words not previously used by Shakespeare.

Thisted and Efron did their analysis for words used up to 100 times before and published it in *Biometrika*, a journal devoted to the theory of statistics. But, to Efron and Thisted, it was just a statistical foray. "It never *possibly* occurred to me that we'd have a chance to use it," Efron remarks, and, in fact, he did not even think of the work when he read reports of Taylor's finding a new poem that may have been written by Shakespeare. Thisted reminded him that the two had done their analysis 10 years ago.

The poem found by Taylor has a total of 430 words from which Efron and Thisted predict it should contain 6.97 ± 2.64 new words. The actual number of new words is 9. (They are admiration, besots, exiles, inflection, joying, scanty, speck, tormentor, and twined.) Seven words of the poem had been used exactly once before by Shakespeare. The prediction was 4.21 ± 2.05. Five words appeared exactly twice before and 3.33 ± 1.83 were predicted. So far, Thisted and Efron have carried their analysis to words used 100 times before and "the poem keeps coming out beautifully, even using quite delicate statistical tests." It looks as though Shakespeare could have written it. Or, to use the more conservative terminology of the statisticians, there is no convincing evidence for rejecting the hypothesis that Shakespeare wrote it.

Efron and Thisted also looked at poems by John Donne, Christopher Marlowe, and

Ben Jonson. None are even close to the predictions of their model for Shakespeare's works. For example, in a poem by John Donne, there were 17 words that Shakespeare never used, although the prediction was for about 8 in a poem of that length.

Persi Diaconis, a statistician at Stanford who is familiar with Efron and Thisted's analysis, says it has altered his own opinion of the newly discovered poem. "I read the poem and it didn't sound anything like Shakespeare to me. I thought any sort of numerical analysis would show that the words are wrongly distributed. But now that I've seen Brad and Ron's analysis, it seems quite plausible to me that the poem could have been written by Shakespeare. I'm as convinced as I would be by any other authenticity check."

Efron stresses that their analysis cannot prove that Shakespeare wrote the poem. But, he says, he is amazed that the newly discovered poem "fits Shakespeare as well as Shakespeare fits Shakespeare." ■

GINA KOLATA

ADDITIONAL READING

R. A. Fisher, A. S. Corbet, C. B. Williams, "The relation between the number of species and the number of individuals in a random sample of an animal population," *J. Anim. Ecol.* **12**, 42 (1943).

B. Efron and R. Thisted, "Estimating the number of unseen species: How many words did Shakespeare know?" *Biometrika* **63**, 435 (1976).

Is Cygnus X-3 a Quark Star?

From a distance of 37,000 light years, the most luminous x-ray source in the galaxy seems to be showering the earth with a new kind of particle; could it be quark matter?

DURING a 10-day period in October 1985, at a time when the galactic x-ray source Cygnus X-3 was undergoing its most violent outburst on record, a flurry of anomalous cosmic ray events from the direction of Cygnus appeared in a proton decay detector deep in Minnesota's Soudan iron mine.

The Minnesota physicists are the first to urge caution: like the events they reported last spring, these October data are inconsistent with any known elementary particle. However, the earlier events have also survived every attempt to explain them away, and the more recent events have markedly improved the clarity of the signal. If the data are real, then the ultrarelativistic debris from Cygnus X-3 contains something totally new to particle physics.

"My gut feeling is that the signals are spurious in some way we haven't understood," says University of Wisconsin theorist Francis Halzen, who has become deeply involved in interpreting the Cygnus phenomenon. "But even if there is only a 1 in 10 chance that they are right, the implications are so important that they must be investigated."

Indeed, if the Soudan events are taken at face value, one of the first implications is that "neutron" stars such as Cygnus X-3 may not be made of neutrons at all. They may instead be spheres of degenerate quark matter.

Cygnus X-3 itself is not a particularly bright source from a terrestrial standpoint; as the name suggests, it is only the third strongest x-ray source in the constellation of Cygnus. On the other hand, it lies some 37,000 light years away, on a far edge of the galaxy where it is heavily obscured by interstellar gas and dust. Intrinsically, Cygnus X-3 is one of the two or three most luminous objects in the galaxy; it and perhaps a few other such sources seem to produce all the ultrahigh–energy cosmic rays above 100 to 1000 trillion electron volts (TeV).

Cygnus X-3 appears to be a binary system consisting of a compact object—call it a neutron star for now—pulling in a stream of gas from a more or less normal companion star; in the process the gas is heated sufficiently to produce the x-rays. The angle of the system is such that the neutron star is eclipsed once every orbit as it passes behind the larger companion. Thus, the corresponding rise and fall of the x-ray signals observed on earth gives a measure of the orbital period: 4.79 hours. However, the source is far from steady. In September 1972, Cygnus X-3 gained astronomical notoriety with an outburst that increased its radio emissions a thousandfold. Since then, smaller outbursts of varying strengths have appeared every 367 days. No one yet understands why the star flares, much less why it does so periodically. Perhaps the normal companion undergoes periodic pulsations of some kind, or perhaps there is a third body that orbits the two companions and regularly perturbs them. But whatever triggers the flares they are exceedingly violent events. During the outburst of October 1982, Ken Johnston of the Naval Research Laboratory was able to detect the shock wave using the Very Large Array near Socorro, New Mexico: it was expanding at roughly one-third the speed of light.

The most recent burst, which lasted from 3 October through 13 October 1985, came at the predicted time within a day and proved to be the largest ever. Observations were made from the ground at radio and infrared wavelengths, and from the European Space Agency's Exosat spacecraft at x-ray wavelengths. Although the astronomers are still reducing and cross-correlating their data, says Johnston, he, for one, is excited. "It's adding a whole new dimension to the model," he says.

What makes Cygnus X-3 a particle physics problem, however, is not the astrophysics but the underground data. The first indications came in 1983, when showers of muons from the general direction of Cygnus X-3 began to show up in the prototype proton decay detector operated in the Soudan mine by physicists from the University of Minnesota and the Argonne National Laboratory. The effect was small: when the Minnesota/Argonne group published its results in the spring of 1985, they only had 60 anomalous events from a 3-degree cone around Cygnus X-3 out of a total background of 1200 events. But those 60 events came with a period of precisely 4.79 hours, and stayed precisely in phase with the radio, x-ray, and infrared emissions. "It's like picking out a lighthouse on a foggy night," says Minnesota's Marvin Marshak.

What made these particular muon showers so striking, aside from their association with an object 37,000 light years away, was that they seemed to have no explanation in terms of known physics. Since muons are unstable and short-lived, they are presumably produced by some kind of primary particle from Cygnus X-3 interacting with the earth's atmosphere or with the rock