Statistics in the Courtroom: United States v. Kristen Gilbert

George W. Cobb and Stephen Gehlbach

George Cobb (GCobb@MtHolyoke.edu) is Robert L. Rooke Professor of Statistics at Mount Holyoke College. He is also currently Vice-President of the American Statistical Association. He received an AB in Russian from Dartmouth, MS in Biometry from the Medical College of Virginia, and PhD in statistics from Harvard. His current interests include statistics education, design of experiments, statistics and the law, Markov Chain Monte Carlo, and DNA microarrays.

Stephen Gehlbach is professor and Dean emeritus of the School of Public Health and Health Sciences at the University of Massachusetts at Amherst. He is a clinical epidemiologist with an MD degree from Case-Western Reserve University and an MPH from the University of North Carolina-Chapel Hill. Recent research activities have involved developing statistical models for predicting mortality in intensive care units, and clinical recognition of osteoporosis and vertebral fracture.

A nurse accused. By the mid-1990s, Kristen Gilbert had been working for several years as a nurse at the Veteran's Administration Hospital in Northampton, Massachusetts. For a time, she had been one of the nurses the others most often looked up to as an example of skill and competence. She had established a reputation for being particularly good in a crisis. If a patient went into cardiac arrest, for example, she was often the first to notice that something was wrong. She would sound the "code blue" or code -- the signal that brought the aid of the resuscitation team. She stayed calm, and she knew how to give a shot of the stimulant epinephrine, a synthetic form of adrenaline, to try to restart a patient's heart. Often the adrenaline did its job, the heart began to beat again, and the patient's life was saved.

Lately, though, other nurses had become increasingly suspicious that something was not right. To some, it seemed that there were too many codes called, too many crises, when Gilbert was on the ward. Over time, the suspicions became stronger. Several patients who went into arrest died, and to some of the staff, the number of deaths was a sinister sign. An investigation was launched. Although an initial report by the VA found that the numbers of deaths were consistent with the patterns at other VA hospitals, the suspicions of the staff remained. Eventually, after additional investigation, including a statistical analysis by one of us (Gehlbach), Assistant US Attorney William Welch convened a grand jury in 1998 to hear the evidence against Gilbert. Welch accused her of having killed several patients by giving them fatal doses of heart stimulant, and he wanted her indicted for multiple murders.

Kristen Gilbert was a mother with two young children. Although she was divorced, she had been dating a male friend for some time. She had a steady job, one that paid reasonably well, and her skill as a nurse was generally recognized. What could possibly motivate her to commit the murders that she was now suspected of? These were not "mercy killings": The victims in the indictment Welch was seeking were not old men or in poor health; rather, they were middle-aged, and healthy enough that their deaths were unexpected. According to prosecutor Welch, Kristen Gilbert did have a reason for what she had done. She liked the thrill of a crisis, she needed the recognition that came from her skillful handling of a cardiac arrest, and, especially, she wanted to impress her boyfriend, who also worked at the hospital.

Part of the evidence against Gilbert dealt with her motivation, part of it came from the testimony of co-workers about her access to the epinephrine she was accused of using in the alleged murders, and part came from a physician who testified about the symptoms of the men who had died. Taken together, this evidence was certainly suggestive, but would it be convincing? No one had seen Gilbert give fatal injections, and although the patients' deaths were unexpected, the symptoms could have been considered consistent with other possible causes of death. It turned out that a major part of the evidence against Gilbert was statistical.

Statistical hypothesis testing I. A key question for the grand jurors was this: Was it true that there were more deaths when Kristen Gilbert was working? Not just one or two extra deaths -- one or two could easily be due just to coincidence -- but enough extra to be truly suspicious? If not, there might not be enough evidence to justify bringing Gilbert to trial. On the other hand, an answer of yes would call for an explanation, and enough other evidence pointed to Gilbert to make an indictment all but certain.

The prosecutors recognized that the key question about excess deaths was one that could only be answered using statistics, and so they asked Stephen Gehlbach, who had done the statistical analysis of the hospital records, to present a summary of the results to the grand jury. In what follows, we will present you with a similar summary of the statistical evidence. As you read through the summary, imagine yourself as one of the grand jurors: do you find the evidence strong enough to bring Gilbert to trial?

The statistical substance involves hypothesis testing, a form of reasoning that uses probability calculations to decide whether or not an observed outcome should be regarded as so very unusual -- so extreme -- that it qualifies as a "scientific surprise." The logic and interpretation of hypothesis testing is fundamental to a lot of work in the natural and social sciences, important enough that anyone serious about understanding how science works should understand this form of reasoning. Unfortunately, in many statistics courses, the logic of hypothesis testing is taught at the same time as some of the probability calculations that you need for particular applications, and the details of the computations tend to eclipse the underlying logic. Part of the challenge facing Gehlbach was to make the logic clear to the grand jury, without going into the details of the calculations.

Gehlbach's testimony to the grand jury. Dr. Gehlbach's testimony was delivered orally, with Gehlbach in the witness stand, talking to the members of the grand jury. The next several paragraphs summarize three parts of Gehlbach's testimony, a first part about the pattern of deaths, by shift and by year, on the medical ward where Gilbert worked, a second part about variability and *p*-values, and a third part about a statistical test for whether the pattern linking the excess deaths to Gilbert's presence on the ward was too extreme to be regarded as due to ordinary, expectable variability. The summaries don't use the exact words from the grand jury testimony, but they cover some of the same substance.

Part One: The pattern of deaths.

(Imagine that at this point in Gehlbach's testimony, the jurors are looking at a graph like the one in Display 1.)

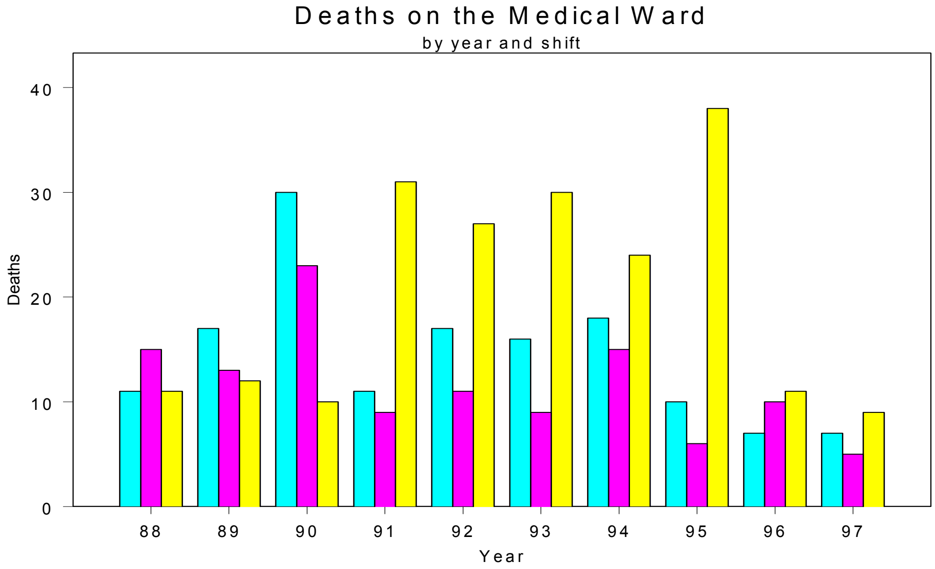
Dr. Gehlbach: "The graph you see shows data from the VA hospital where Kristen Gilbert worked. Each set of three bars shows one year's worth of data, starting in 1988, and running through 1997. Within each set of three bars, there is one bar for each shift. The left bar is for the night shift, midnight to 8 a.m.; the middle bar is for the day shift, 8 a.m. to 4 p.m., and the right bar is for the evening shift, 4 p.m. to midnight. The height of each bar tells how many deaths there were on a shift for the year in question.

"Now look at the pattern from one year to the next. For the first two years, '88 and '89, the bars are short, showing roughly ten deaths per year on each shift. Then there is a dramatic increase. For the years 1990 through 1995, there is one shift in each set of three with 25 to 35 deaths per year. Then for the last two years, the bars are all short again, a bit under ten deaths per year on each shift.

"How does this pattern fit with Kristen Gilbert's time at the VA? It turns out that Ms. Gilbert began work on Ward C, the medical ward, in March of 1990, and stopped working at the VA in February of 1996. Looking at the deaths by year, the pattern tracks Ms. Gilbert's work history: small numbers of deaths in years when she didn't work at the VA, and large numbers when she was there.

"We can learn more by looking at the different shifts. You'll notice that in each of the years that Ms. Gilbert worked on Ward C, one of the three shifts always shows more deaths than the other two. For five of these six years, 1991 through 1995, it's the evening shift that stands out. During these five years, Ms. Gilbert was assigned to the evening shift.

"What about the exception, 1990? That year it is the night shift, not the evening shift, that stands out as having an unusually large number of deaths. Well, it turns out that 1990 was also an exception for Kristen Gilbert's work history. That year she was assigned not to the evening shift, but to the night shift.



Display 1: The pattern of deaths, by year and by shift
In each set of three bars, left = night (midnight - 8 am), middle = day (8am - 4 pm), right = evening (4pm - midnight).

At this point in the argument, there is a clear pattern associating Gilbert's presence with excess deaths. However, in principle the pattern *might* be nothing more than the result of ordinary, expectable variation. The goal of a statistical test in this situation would be to determine whether the numbers of excess deaths were too extreme to be accounted for by such variation. In order to prepare the jurors to think about a statistical test here, Gehlbach first explained the basic ideas in a more familiar context:

Part Two: Variability and p-values.

Dr. Gehlbach: "To understand the idea of a statistical test, think about tossing a coin. How can you decide whether there's something suspicious about a set of ten coin flips? Ordinarily, we expect a coin to be fair, which means there's a 50-50 chance of heads. This is our hypothesis, the starting point of our reasoning. If you flip ten times, and the coin is fair, then on average you'd expect five heads to show up. But you know that you might get six, seven, or even eight heads. Things vary, and sometimes the variation is due just to chance.

"Now suppose you got ten heads in ten flips. Is that result extreme enough to be suspicious? How extreme an outcome do we need before we should doubt our hypothesis that the chance of heads is 50-50?

"To answer this question, statisticians compute a *p-value*: start with the hypothesis – a 50-50 chance of heads – and compute the probability of six heads, or seven heads, and so on. It turns out that the probability of at least six heads in ten flips is about 0.38. This means that 38% of the time when you make ten flips with a fair coin, you'll get at least six heads. If something happens 38% of the time, there's nothing surprising or suspicious about it.

"For seven heads, the p-value is about 0.17: About 17% of the time you do ten flips of a fair coin, you'll get seven heads or more. So seven out of ten isn't really surprising either.

"If you get nine heads in ten flips, however, that's unusual. If the coin is fair, then you're unlikely to get a result that extreme. The probability, or p-value, for 9 or more heads, is only about 0.01, or 1%.

"For ten out of ten, the p-value is about 0.001, or one in a thousand. This is a result so extreme that you'd almost never get it from a fair coin. If you saw me pull a coin out of my pocket, flip it ten times, and get heads every time, you'd be justified in thinking there's something going on besides just chance variation.

"That's how statisticians use a p-value. If we see a result with a really low p-value, then either we've seen a really rare outcome, or else the hypothesis we used to compute the p-value must be wrong.

"In many medical trials -- testing whether antihistamines relieve your symptoms of allergies, or things like that – we compute a p-value assuming the drug has no effect, and a probability of one out of a hundred is unusual enough, and the evidence would be considered strong enough, to conclude that the medicine actually worked.

Now, with the basic logic out on the table, it was time to present a formal test. What follows is just one very focused part of the set of tests Gehlbach actually presented.

Part Three: A statistical test.

At this point Gehlbach showed the jury data like the table in Display 2.

Dr. Gehlbach: "The table summarizes records for the eighteen months leading up to the end of February, 1996. (That February was the month when Ms. Gilbert's co-workers met with their supervisor to express their concerns; shortly after that, Ms. Gilbert took a medical leave.) With 547 days during the period in question, and three shifts per day, there were 1641 shifts in all. Out of these 1641 shifts, there were 74 for which there was at least one death.

"Now think of each shift as like a coin flip, with a death on the shift if the coin lands heads. The fraction of shifts with a death is 74/1641 or 0.045. This means that out of every 100 shifts, you would expect four-and-one-half shifts, or 4.5%, with at least one death. It's like tossing a coin, one toss per shift, with a probability of 0.045 that the toss lands heads.

> "Now let's look just at the shifts when Ms. Gilbert worked. There were 257 of these. If the deaths distributed themselves like coins landing heads, we'd expect between 11 and 12 of these shifts to experience a death, because 4.5% of 257 is 11.6.
>
> "What does the record show? As you can see from the table, there were in fact not 11 or 12 shifts with a death, but 40. How extreme is 40? Could you get a number like that just from chance variation, or is 40 really suspicious? To answer that, we compute a p-value.
>
> "Assume that the 257 shifts that Ms. Gilbert worked behaved like coin tosses, with a chance of heads equal to 0.045. What is the probability of 40 or more deaths? The p-value turns out to be less than one in one hundred million. In other words, it is virtually impossible to get as many as 40 shifts with deaths from ordinary, chance-like variation.

Number of days	547
Number of shifts	1641
Number of deaths	74
Deaths per shift	0.045
Shifts with KG present	257
Expected number of deaths	11.59
Observed number of deaths	40

Gilbert	Death on Shift?		
present?	Yes	No	Total
Yes	40	217	257
No	34	1350	1384
Total	74	1567	1641

Display 2: The basis of the statistical test

Gilbert on trial. The grand jury found the evidence persuasive and Gilbert was indicted. Because the VA hospital is legally the property of the federal government, Gilbert would stand trial in federal district court, on four counts of murder and three additional counts of attempted murder. The question of jurisdiction was important because, although the state of Massachusetts has no death penalty, Gilbert was facing a federal indictment, governed by federal rather than state laws, and US Attorney Welch decided to ask for the death penalty. Kristen Gilbert would be on trial for her life.

Before the trial got under way, the judge, Michael A. Ponser, had to rule on whether the jury should be allowed to hear the statistical evidence. On the one hand this seems like a "no-brainer." After all if the evidence was an important part of what was presented to the grand jury, if it was appropriate for them to hear, and they found it compelling, what could possibly be wrong with letting the trial jury hear the same testimony? On the other hand, a counterargument might be that allowing the statistical evidence would just lead to the unhelpful distraction of "dueling experts." The court system allows expert testimony when the evidence involves specialized technical or scientific issues that go beyond what members of the jury would ordinarily be familiar with. The purpose of the experts is to provide explanations of the science or of the technical facts involved, along with the appropriate conclusions -- in other words, to help the jury understand the evidence better -- and the US Supreme Court has set guidelines aimed at making sure that unscientific testimony is not admitted. The goal is to help ensure that the verdict will be scientifically sound. Nevertheless, attorneys sometimes say that if there is expert testimony on one side, the other side hires another expert who will disagree, and the jury, rather than think through the explanations, will simply ignore it all. One expert cancels the other. Although this view may be overly cynical, no doubt it does have a basis in fact.

Rather than just rely on the crude strategy of "dueling experts," Gilbert's defense attorneys asked the other of the two of us (Cobb), to prepare a written report for the judge summarizing the reasons why it would not be appropriate for the new jury, the trial jury, to hear the same evidence that Gehlbach had presented earlier to the grand jury. In the next several paragraphs, you will read a summary of the main points in that report. This time, put yourself in the position of Judge Ponsor: Do you find these points persuasive? Would you have allowed the jury to hear the statistical evidence, or not?

Hypothesis testing II. So far, in the Gehlbach testimony, the interpretation of hypothesis testing has focused on what it is that a tiny *p*-value *does* tell you. It tells you that the observed result is too extreme to be explained as due to chance-like variation. This was exactly the relevant issue for the grand jury: Were there so many excess deaths when Gilbert was present as to be suspicious in the eyes of science? The clear answer was yes. In the Cobb report, the focus was on things that tiny *p*-values do *not* tell you. Unfortunately for people who need to understand hypothesis testing, these invalid conclusions are a constant temptation. They seem to make sense intuitively, but they are wrong, and so they have great potential to mislead the unwary. This potential for logical mischief was the basis for the defense team's request that Judge Ponsor not allow Attorney Welch to present statistical evidence to the trial jury.

Cobb's report to Judge Ponsor. Leaving aside a variety of secondary technical issues, the Cobb report made three main points. One of them was to agree with the bottom line conclusion in Gehlbach's testimony. The other two dealt with two limitations on what you can learn from a tiny *p*-value.

Point One: The defense and prosecution statisticians agree! As mentioned earlier, often the two experts who provide testimony on scientific evidence disagree. However, that was *not* what happened in the Gilbert case. Cobb's report *agreed* with Gehlbach's testimony before the grand jury. We both thought the pattern linking Gilbert's presence on the ward with excess deaths was far, far too strong to be regarded as mere coincidence due to chance-like variation. We both thought, too, that in the absence of any innocent explanation for the pattern, the association was more than strong enough to justify the indictment. Why then, shouldn't the trial jury hear the testimony? To answer that question we proceed with Cobb's other two points.

Point Two: Association is not causation. The grand jury and the trial jury have quite different decisions to make as they weigh the evidence, and the difference is closely tied to what a *p*-value does and does not tell you. The grand jury had to decide whether or not Gilbert should stand trial. Was there enough suspicion to justify the expense to the government and the psychological burden on Gilbert to hold what promised to be a long and expensive trial? A grand jury does not have to decide guilt or innocence beyond a reasonable doubt. For them, the standard is much lower. They are simply asked to determine whether the level of suspicion is high enough. This is precisely the kind of question that logic of hypothesis testing is designed to answer. In statistics, and in science generally, the bar is set quite high for what deserves to be considered strong suspicion, typically a *p*-value of .05 or .01.A low *p*-value establishes suspicion by *ruling out chance variation* as an explanation. Notice that a low *p*-value does not *provide* an explanation. It doesn't say, "Here. This is the reason for the excess deaths." What it says is much more limited: "Whatever the explanation may be, you can be quite confident that it is *not* mere chance variation."

The trial jury isn't asked to decide whether the facts look suspicious. By the time a case comes to trial, the decision about suspicion has already been made. The trial jury is asked to decide the reason for the suspicious facts. Were the excess deaths caused by Gilbert giving fatal injections? Or were there enough uncertainties that the cause could not be determined beyond a reasonable doubt? Because a low *p*-value cannot tell you about cause, the Cobb report argued, the statistical evidence was not an appropriate part of the evidence for the trial jury.

But wait. Isn't statistical evidence used all the time to draw conclusions about cause? Doesn't the FDA use statistics to decide whether a particular medication will cause a disease to go away, or at least cause its symptoms to go away, or whether that same medication will cause side effects? Didn't scientists use hypothesis testing to decide, for example, that antihistamines can relieve the symptoms of allergies? If hypothesis testing can tell us about cause in these situations, why not in the Gilbert case also?

The answer involves what some statisticians consider to be the single most important contribution that statistics has made in the last 100 years: an understanding of the difference between an observational study and a randomized experiment. The statistical analysis in the Gilbert case is based on observational data; in the studies used to decide such things as whether taking aspirin lowers the risk of heart attacks, the data come from randomized experiments. The distinction here is so important that it is worth pausing to take a look at it in more detail.

An important distinction: Observational studies versus randomized experiments. A famous study from the early research on smoking and health illustrate why observational studies can be misleading about cause and effect. Look at the death rates from that study, and notice what the "obvious" conclusion would be:

Non-smokers	20.2
Cigarette smokers	20.5
Cigar and pipe smokers	35.3

Display 3: Death rates, in deaths per 1000 people per year

The sample sizes in this study were huge, and the tiny *p*-values conclusively rule out chance variation as an explanation for the differences among the three groups.

Taking the numbers at face value would leave us with the conclusion that cigarette smoking carries only a miniscule risk, but that pipes and cigars are highly dangerous.

To avoid this logical trap, you need to recognize that a low *p*-value, by itself, does not prove a cause-and-effect relationship; it only eliminates chance as one of the possible cause. For this study, there was another cause at work behind the scenes: age. The non-smokers, on average, were 54.9 years old; the cigarette smokers only 50.5, and the cigar and pipe smokers were 15 years older, at 65.9. Because the researchers had this additional information, they were able to use statistical methods to adjust for the effect of the “lurking variable,” age. The adjusted death rates are in line with what we have come to expect:

Non-smokers	20.3
Cigarette smokers	28.3
Cigar and pipe smokers	21.2

Display 4: Death rates, adjusted for age, in deaths per 1000 people per year

For our purposes, the key point is this: With an observational study, you can never know for certain whether your numbers look the way they do for the reasons you know about, or whether, instead, there are hidden causes at work. With a randomized experiment, the groups being compared are created using randomness or chance. If the group sizes are large enough, the randomization process evens out all possible influences that might make one group different from another. The beauty and power of the randomization is that it evens out all unwanted influences, including the ones you don’t know about.

Consider how randomization worked for an influential study of aspirin and heart attacks. Back in the 1980s, researchers began a huge study involving 21,996 physicians across the US. All of them had volunteered to take part. Some were older; some were younger. Some were overweight; others were not. Some exercised regularly; others didn’t. Cholesterol levels varied from low to very high. In short, there were many influences, both known and unknown, that caused big differences in the risk of heart attack. To ensure that all these influences would even out, the researchers used a chance device to assign each physician to one of two groups. Those assigned to the treatment group took a daily pill that actually contained aspirin. Those assigned to the placebo group also took a daily pill, one that was identical to the other pill, except that it contained no aspirin.

The results of the study were striking. Even before the study was supposed to have ended, preliminary *p*-values were so low that the scientists in charge of the research decided to call an end to the experiment so that the physicians in the placebo group could start taking the aspirin if they wanted to.

In a nutshell, here is the key difference between the experiment and the observational study. In the observational study, when a low *p*-value ruled out chance variation as an explanation for the differences in death rates, it was not clear what was causing

> those rates to differ. The difference in rates might have been caused by differences in smoking habits, might have been caused by differences in age, and might have been caused by any of a number of other influences. The *p*-value alone was of no help in deciding the cause. In the experiment, one possible cause -- aspirin -- was singled out for investigation, and the experiment was carefully designed to eliminate all other possible causes, apart from chance-like variability. When a statistical test was then able to rule out chance variability, there was only one possible explanation left, the cause that was singled out for study. Observational studies can be useful; after all they did play a large part in making the link between smoking and cancer, but one must work very hard to eliminate the many uncontrolled factors as possible explanations for the observed association. Randomized experiments make it much easier to draw causal conclusions.

The data in the Gilbert trial was observational. To make it an experiment, Gilbert's presence on the ward would have had to be assigned using a chance device to decide which shifts she worked. Because there was no experiment, the tiny *p*-value, though it ruled out chance as an explanation for the excess deaths, did not rule out other possible explanations. In his ruling, Judge Ponsor gave a hypothetical example that would produce similarly damaging statistical evidence: Suppose that on a shift when Gilbert was present, a boiler had accidentally exploded and killed several dozen patients. Such an accident could lead to a tiny *p*-value showing that Gilbert's presence was associated with a very high number of deaths, but it would not be evidence of her guilt. As statisticians often say, "Association is not causation." The temptation, especially when the evidence of association is strong and there is a plausible explanation, is to conclude that the test provides evidence that the explanation is right. With observational data, such a conclusion would be based on false logic.

Point Three: The Prosecutor's Fallacy. The Cobb report pointed out a second, closely related temptation that is present with hypothesis testing. The *p*-value is a conditional probability, computed by assuming that a result is due to chance-like variation. It summarizes logic that goes as follows: "If the cause is just random variation, then the extreme result is very unlikely. We got an extreme result. Therefore, it is not reasonable to think that random variation is the cause." Notice that this logic says nothing about other causes. If there was a boiler explosion, for example, the extreme result would not be at all surprising.

Now look at how slippery the logic can get if you're not careful: "Suppose Gilbert is not guilty, and that the deaths behave in a chance-like way, like coin tosses. Then the probability is less than 1 out of a hundred million that you would see so may excess deaths on Gilbert's shifts." (Correct.) It's a quick jump to the following shorter version: "If Gilbert is innocent, then it would be almost impossible to get so many excess deaths." (Also correct.) And then, "With this many excess deaths, the chance is less than 1 in a hundred million that Gilbert is innocent." (*Not* valid.) This kind of "reasoning" is so tempting, and so common, that it has become known to statisticians as the prosecutor's fallacy. Because the false logic beckons so seductively, it is often used as the basis for arguing, as the Cobb report did, that the statistical evidence was likely to be misinterpreted by the jury in a way that favored the prosecution, and was therefore “prejudicial.”

Conclusion. Judge Ponsor ruled that the statistical evidence should not be allowed at trial. Nevertheless, the other, non-statistical evidence proved to be enough to convince the jury, and, after many days of deliberation, Gilbert was convicted on three counts of first degree murder, one count of second degree murder, and two counts of attempted murder. After a penalty phase of the trial, the jury voted 8 to 4 for a death sentence, and because the vote was not unanimous, Gilbert’s life was spared. She is now serving a sentence of life in prison without possibility of parole.

The statistical analysis that uncovered the pattern linking Gilbert’s presence to the excess deaths was an essential part of the process that brought her to justice. The two juries that Gilbert faced, and their different roles in our system of justice, illustrate neatly the proper interpretation of hypothesis testing. First, a small *p*-value *does* allow you to rule out chance-like variability as a plausible explanation for an observed pattern. It tells you that that the observed pattern is so extreme as to qualify as a surprise in the eyes of science. Second, if your data are observational, a small *p*-value does *not* tell you what has *caused* the surprise. Association is not causation. Inferences about cause are much more straightforward with a randomized experiment.

Additional Readings:

Cameron, J.B. (2001), “Gilbert guilty of four murders,” *Daily Hampshire Gazette*, March 15, 2001.

DeGroot, M. H., Fienberg, S. E., and Kadane, J. B. (1994), *Statistics and the Law*. John Wiley.

Finkelstein, M. O. (2001), *Statistics for Lawyers, 2nd edition*. Springer-Verlag.

Gastwirth, J. L. (Ed.). (2000), *Statistical Science in the Courtroom*, Springer- Verlag.

Gastwirth, J. L. (1998), *Statistical Reasoning in Law and Public Policy*, Academic Press.

Good, P. I. (2001), *Applying Statistics in the Courtroom*, CRC Press.

Zeisel, H., and Kaye, D. H. (1997), *Prove it with Figures*, Springer-Verlag.