COLUMBIA LAW REVIEW
SIDEBAR

LAW, STATISTICS, AND THE REFERENCE CLASS PROBLEM

Edward K. Cheng[*]

<u>Preview of</u>: Edward K. Cheng, A Practical Solution to the Reference Class Problem, 109 Colum. L. Rev. (forthcoming Dec. 2009).

Statistical data are powerful, if not crucial, pieces of evidence in the courtroom. Whether one is trying to demonstrate the rarity of a DNA profile, estimate the value of damaged property, or determine the likelihood that a criminal defendant will recidivate, statistics often have an important role to play. Statistics, however, raise a number of serious challenges for the legal system, including concerns that they are difficult to understand, are given too much deference from juries, or are easily manipulated by the parties' experts. In this preview piece, I address one of these challenges, known as the "reference class problem," and sketch a solution that I develop at greater length in my forthcoming Essay.[1]

I. THE REFERENCE CLASS PROBLEM

The reference class problem arises from a basic observation: When we make statistical inferences about a specific case, those inferences depend critically on how we group or classify that case. To illustrate, imagine that plaintiff contracts cancer after being exposed to a chemical spill of a known carcinogen. To establish that the spill is the cause of her cancer, plaintiff attempts to show that her cancer risk doubled after exposure.[2] So far, the litigation seems pretty straightforward, but then

---

1. Edward K. Cheng, A Practical Solution to the Reference Class Problem, 109 Colum. L. Rev. (forthcoming Dec. 2009).

2. See 3 David L. Faigman et al., Modern Scientific Evidence § 23:27, at 249 (2008) ("Most courts . . . have concluded a plaintiff can reach a jury if she can present epidemiological studies indicating at least a doubling of the risk of injury due to exposure to a substance . . . .").

we face a dilemma. What statistic should we use to estimate plaintiff's cancer risk? Should we use the risk for the general population, or should we be more specific? White females under the age of fifty? Residents of Littleton County with no family history of cancer? In other words, in describing cancer risk, how should we break down the population: by age, gender, geography, profession, or something else?

In any litigation, parties will invariably offer different classifications in the hope of gaining some advantage. To minimize her background risk, our plaintiff may suggest using women under the age of fifty with no family history of cancer as the relevant group. In contrast, the defendant will focus on other attributes, such as the fact that she is a smoker or takes hormone supplements. Faced with these conflicting statistics, what is a jury to do? One natural response is to use all of the information about the plaintiff—but that would result in a class of one person, the plaintiff herself, and that singular class does not enable us to make any statistical inferences at all.

The reference class problem thus presents a serious issue. The use of statistics is supposed to increase objectivity and rigor, yet as I describe it above, statistics appear almost infinitely malleable: As long as counsel manipulates the reference class sufficiently, he can arrive at any background risk number he wants. Indeed, rather than fulfill their promise as a neutral basis for decisionmaking, statistics suddenly appear to be nothing but rhetorical tricks that advocates can deploy in court.

Worse yet, this problem is not confined to toxic tort cases; it arguably infects every use of statistics in the law. For example, when courts value property for eminent domain, taxation, or insurance, one standard method is to look at comparable properties. But which properties are in fact "comparable" and what attributes of a home or lot should be used for the valuation? The choice of reference class can affect the valuation considerably. In DNA cases, prosecutors often emphasize the random match probability (RMP), the probability that a person chosen at random from the population will have the same profile as the one found at the crime scene. Yet, what population is appropriate for calculating the RMP? The entire human population? The defendant's racial subgroup? The city in which the crime occurred?

From an intuitive standpoint, the above discussion may seem somewhat alarmist. After all, just because one can manipulate statistical inferences by cleverly selecting reference classes does not necessarily mean that a jury will buy them. Using the category of white females under age fifty to estimate cancer risk seems natural and hence legitimate. Using the category "women who own blue handbags, like sushi, and drive a red sedan" does not. But relying on the jury's powers of intuition carries two problems. First, mindful of the jury's skepticism, the parties will never offer outrageous reference classes. They will instead choose plausible (but still conflicting) ones to advance their case. Under these conditions, the jury's intuitive judgment is largely unhelpful, and its choices effectively arbitrary. Second, to rely solely on

intuition is to surrender the goal of using statistics to inject greater objectivity and rigor into legal decisionmaking. As Ron Allen and Mike Pardo recently noted, if reference class selection ultimately boils down to subjective and intuitive judgment, then statistical models of evidence have not advanced the field by much.[3]

But what if we could find a way to make this intuition about the "reasonableness" of reference classes more rigorous? Providing a principled method for choosing one reference class over another would arguably solve the reference class problem, or at least restrict its potential for mischief. To add this rigor, my proposal draws a close analogy to the model selection problem in statistics and applies those concepts and methods to the reference class problem.

II. MODEL SELECTION

A straightforward way to understand model selection is to consider the problem of fitting a line or curve to a set of points.[4] For example, assume we would like to predict a student's GPA based on the number of hours he/she studies. We collect the data shown in Figure 1a, and then the question becomes, what exactly is the relationship? The most obvious answer is a simple linear relationship, as in Figure 1b. However, the slight curve in the data points might suggest a quadratic relationship, as in Figure 1c. We can fit even more complex curves, such as the fourth degree polynomial in Figure 1d. In any event, we have multiple candidates for models and no obvious principle for choosing one over another.

We can of course select curves based on intuitive judgment. For example, the fitted curve in Figure 1d is obviously overcomplex: Study hours and GPA are unlikely to be related in this way. Indeed, this kind of intuition may be what underlies the time-honored principle of Occam's Razor.[5] But intuition does not tell us how or why the curve is *excessively* or *unnecessarily* complex. Intuition is neither precise nor objective. It can exclude the fourth degree model with ease, but has a harder time choosing between the linear and quadratic curves.

The statistics literature, however, does offer a more rigorous perspective on the model selection problem. Complex models like Figure 1d are problematic because they are "overfitted." The problem with overfitted models is that they erroneously incorporate the random noise that accompanies real world data. As a result, the predictions they

---

3. See Ronald J. Allen & Michael S. Pardo, The Problematic Value of Mathematical Models of Evidence, 36 J. Legal Stud. 107, 115 (2007) ("[T]he question of which [reference class is better will] . . . be the subject of argument and, ultimately, judgment.").

4. See generally Walter Zucchini, An Introduction to Model Selection, 44 J. Mathematical Psychol. 41 (2000) (offering short and less technical introduction to concepts in model selection).

5. See, e.g., Lewis S. Feuer, The Principle of Simplicity, 24 Phil. Sci. 109, 109 (1957) (**"**Entities are not to be multiplied unnecessarily." (emphasis omitted)).

make become less accurate than if they had simply ignored the noise. In the GPA example, if presented with a new set of students and their study hours, the overfitted model will make more errors in predicting GPA than a simpler model. So we have a classic tradeoff. Too simple a model, and it will fail to identify the underlying relationship and have low accuracy. Too complex, and it will incorporate too much random noise and be similarly inaccurate.
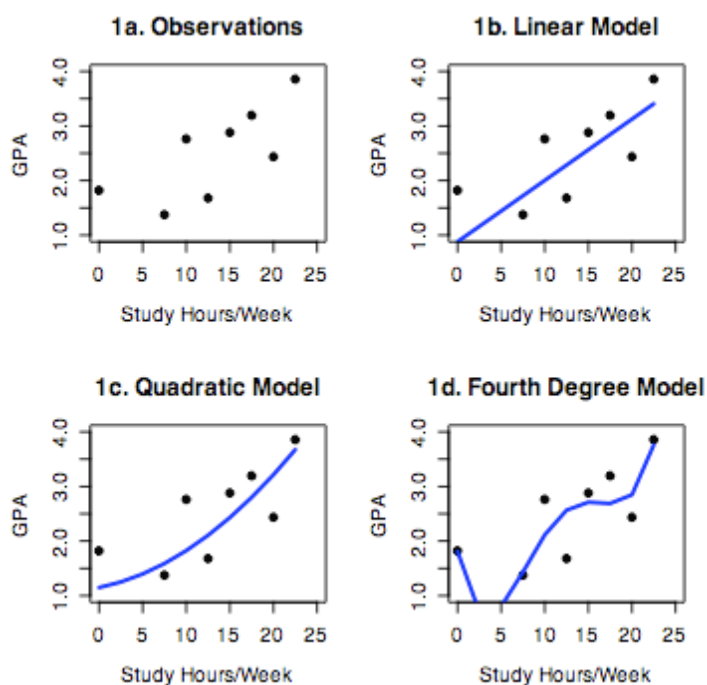

Figure 1: Example Fits to Observed Data Points

To perform this balancing between fit and complexity, statisticians have developed various model selection criteria.[6] These criteria operate as rating systems, allowing researchers to compare different models and select the "best" one.

III. A SOLUTION

At this point, the deep parallels between model selection and the reference class problem are probably evident. Overly narrow reference classes are essentially overly complex models—they take into account too many attributes and run the risk of incorporating noise into their

---

6. See Kenneth P. Burnham & David R. Anderson, Model Selection and Multimodel Inference 31–37 (2d ed. 2002) (discussing the need to balance fit and complexity and various model selection methods).

estimates or predictions. Conversely, overly broad reference classes are like underfit models—they fail to incorporate enough of the information in the data.

Indeed, as I argue in the Essay, the reference class and model selection problems are precisely one and the same. As a result, model selection criteria can solve the reference class problem for all practical purposes in legal proceedings. Choosing a reference class need not be arbitrary, subjective, or intuitive, but rather can be relatively objective and quantifiable. Juries do have principles for selecting which statistics to use to estimate a plaintiff's background risk, a house's market value, or a DNA profile's random match probability.

Predictably, this claim is subject to a number of limitations, the most important of which is that the proposed solution only eliminates the reference class problem in the *legal* context. The reason is that no one has yet figured out how to find the single best model for a given phenomenon. (That problem is exceptionally difficult, if not impossible, to solve.) But as lawyers, we do not need to find the absolute best reference class to resolve issues in court. The adversarial system only requires courts or juries to mediate disputes between the parties, so they just need to decide whose proposed reference class is *better*. And model selection criteria perform that comparative function handily.

CONCLUSION

Beyond the academic aspects of the proposed solution, my hope is that this project will alert practitioners and courts to two fundamental things. For practitioners, as one philosopher of science aptly said, "the reference class problem is your problem too."[7] Whenever you encounter a statistic, think deeply about the underlying reference class. Changing the reference class may change the statistic, and thus allow you to challenge your opponent, make a powerful rhetorical argument, or in the best case scenario, affect the outcome. For courts, the lesson is that the reference class problem is not as intractable as it first seems. The choice of reference class need not be left entirely to a jury's subjective or intuitive judgment. Rather, statistical tools exist for making reference class selection more analytical, a development that will hopefully make statistics more welcome in the future.

Preferred Citation: Edward K. Cheng, *Law, Statistics, and the Reference Class Problem*, 109 COLUM. L. REV. SIDEBAR 92 (2009), http://www.columbialawreview.org/Sidebar/volume/109/92_Cheng.pdf.

---

7. Alan Hájek, The Reference Class Problem Is Your Problem Too, 156 Synthese 563 (2007).