Justice by the Numbers: Educating Judicial Decision Makers

Gray, Mary W. and Al-Shara, Nawar
American University, Department of Mathematics and Statistics
4400 Massachusetts Avenue NW
Washington DC 20016-8050, USA
mgray@american.edu

1. Introduction

In employment cases, in criminal cases, in insurance cases, in antitrust cases, the testimony of statistical experts has become increasingly important, not only in United States courts, but elsewhere in the world. However, generally judges have great discretion in deciding what evidence can be admitted and ultimately the validity and weight of such evidence. Given that not only is statistics not well understood by the general public, but in fact may produce anxiety, sound judicial decision making is in real jeopardy if the decision makers cannot be made to understand and evaluate statistical evidence properly.

2. Admission of evidence

Admission of evidence in U.S. federal courts is governed by the Federal Rules of Evidence, which generally evince a preference for admitting expert evidence, requiring primarily that the evidence be "relevant," which is defined to mean that it "tends to make existence of a consequential fact more or less probable" than it otherwise might have been. In the past, judges operated under the *Frye* rule (1923), which stated that "the things from which the deduction is made must be sufficiently established to have gained general acceptance in the particular field in which it belongs." "Gained general acceptance" in turn usually meant in practice "published in a peer-review journal." Certification by independent experts in the field would, it was reasoned, eliminate "junk" science and assure that a professional standard was meant. With the opinion of the U.S. Supreme Court in *Daubert v. Dow Pharmaceutical* (1993), the standard has changed, placing the burden of deciding whether to hear the evidence squarely on the judge. Under *Daubert* a judge may now decide whether (s)he finds the particular statistical technique worthy of admission with only the guidance of a checklist, offered by the Court with the *caveat* that none of its items is to be considered determinative and that in its entirety it cannot be deemed definitive:

a) whether the theory or technique can be, and has been, tested;
b) whether the technique has been published or subjected to peer review;
c) whether actual or potential error rates have been considered;
d) whether the technique is widely accepted within the relevant scientific community.

In making this determination, the judge needs to evaluate the methodology used, but, according to the recently decided *General Electric v. Joiner* (1997), may also assess the conclusions drawn by the expert, for the two "are not entirely distinct." The *Joiner* court went on to say that "[N]othing ... requires a district court to admit opinion evidence which is connected to existing data only by the ipse dixit of the expert. ... A court may conclude that there is simply too great an analytical gap between the data and the opinion proffered."

In the *Joiner* case the question was whether studies connecting stomach cancer in mice with polychlorinated biphenyls (PCBs) were relevant to whether PCBs might have caused lung cancer in the plaintiff, not really a statistical question although the validity of the meta-analysis technique of the study which was not admitted was also called into question. Although sometimes the question of the admission of statistical evidence hinges on whether the methodology is appropriate, frequently the relevance issue with statistical evidence is of a more fundamental nature. Judges, even if they admit the evidence, may in determining the weight it should be given assert that statistics can have no relevance to an individual case. Thus, in a series of death penalty cases starting with *McCleskey*

(1987), courts have dismissed evidence of systemic bias in sentencing–with a death penalty four times more likely to be given to one convicted of killing a white person than one convicted of killing a black person–not because of doubts about the methodology or the conclusions drawn by the experts, but because they have decided that there must be evidence not merely of a pattern of bias but of actual bias in the case at hand.

2. Methodology

When faced with conflicting conclusions of experts, judges have often thrown up their hands, as, for example, when they have been unable to sort out the merits of rival regression models of salaries of college faculty (*Presseisen* 1978; *Ottaviani* 1989; see also Gray 1993). However, in general, statistical evidence is more likely to encounter judicial resistance if the technique appears untried or esoteric. Keeping in mind that the question of the acceptability of a technique as a guideline for admissibility has not necessarily disappeared with the advent of *Daubert*, the expert will generally find that a journal is a better forum for the introduction of radically new methodology than the courtroom. Unfortunately, even general acceptance by the statistical community may not guarantee judicial acceptance; getting a technique admitted for the first time can present a formidable obstacle. For example, the court in *Wyche v. Marine Midland Bank* (1997) rejected a proportional hazards model on the grounds that no court had previously permitted the use of such a methodology. The judge cited Herbert and Shelton (1996)for the inappropriateness of the model. The issue being addressed in *Wyche* was the proper model to use in age discrimination cases. Federal law in the United States protects only those forty or older, so the question is whether to use a dichotomous or continuous model; the latter can, of course, detect whether discrimination becomes more pronounced as workers get older. Certainly a proportional hazards model can address this issue, but the sophistication of the technique may well have caused its rejection. In *Mistretta v. Sandia Corporation* (1977), plaintiff's expert successfully used a Kolmogorov-Smirnov test–arguably simpler than a proportional hazards model--to show that workers aged 52 to 64 were adversely affected by a layoff although the over-40 population as a whole was not. It should be noted that there are also problems with using a continuous variable to model potential age discrimination because one needs artificially to account for the fact that for evidentiary purposes everyone under forty must be treated in the same undifferentiated way.

In spite of arguments about particular models that may be adopted, regression methodology has been generally accepted by the courts as have, in addition to the ubiquitous *t*-tests, Wilcoxon, Mann-Whitney, Mantel-Haenzel and urn techniques. What has generally not been well received is any technique which, in the view of the judge, hints at biased assumptions. Thus, judges do not like one-tailed *t*-tests–which they often consider as prejudging the existence of disparities–nor do they like any Bayesian techniques, whose proponents have generally failed to make clear to the court how and why the priors are established. The judges fear that somehow the evidence is only "proving" what has already been assumed to be true in setting up the statistical model.

3. Educating the decision makers

The rationale for the role of judges in deciding whether evidence is to be admitted in cases in which the jury is the trier-of-fact is that the judges can determine whether the likelihood of confusion or bias trumps the relevance of the evidence and thus leads to its exclusion. On the other hand, one may ask whether a judge is any better able to screen out "junk" statistics, whether methodologically or otherwise flawed, than is a jury, Although there are ongoing efforts to give judges training in technical fields, in general the level of expertise among the judiciary is not necessarily higher than among the jury population. In fact, most juries know they do not know much about statistics as indeed they know little about many aspects of what is being presented to them, but they also recognize their responsibility to be instructed. Judges, on the other hand, all too frequently are so accustomed to being the authority figure that they do not want to be instructed, particularly about a subject with which they have never felt comfortable. In a concurring opinion in *Joiner*, Justice Stephen G. Breyer suggested that judges rely more on scientific panels to help sort out evidence. Such proposals have been around for a long time, but only rarely have judges used either panels or court-appointed experts to help with the evaluation of technical evidence. However, recently a judicially-appointed panel of experts found no

correlation between silicon implants and a variety of illnesses, thus capturing considerable public attention and substantially influencing the course of litigation. A recent *Washington Post* op-ed article cited approvingly the use of the panel as a return to the original practice of fourteenth century English courts in using expert witnesses (Griffin 1999).

How then to present statistical evidence in a way to get it admitted in the first place and to have it properly understood in the second? Often statistical preferences are sacrificed for the ability to communicate clearly. For example, if one is studying whether men and women who have the same qualifications and are doing the same work are being paid the same salaries, it might seem appropriate to formulate a regression model based on the salaries of the men alone and then to look at the salaries predicted for the women under this model. The average residual for women might then be taken as a measure of the sex-based disparity; this might be refined by modeling the men's salaries on those of the women and comparing their average residual to that of the women under the men's model. But what are the chances of the judge or jury following this analysis? On the other hand, if one simply throws all the men and women into the model, including a variable for sex along the way, the sex coefficient is a nice, convenient way to show the disparity, if any. That is why that is the technique generally used in sex discrimination cases; similarly a model using the log of salaries may be rejected, even though "better" in some sense, because of the complexity it adds, not to mention the frequent judicial confusion of the use of logs with logistic regression to the detriment of understanding and sound decision making (Gray 1993).

Simplicity needs to be balanced with completeness, theoretical purity with practical application. It is important to note that while experts are generally accorded wide latitude in their testimony–only they are allowed to express opinions and to respond to hypothetical questions–nonetheless the evidence not only ought to have a substantial effect on the probability of the existence of a material fact but it should be apparent *why* that is the case. In a landmark cases involving charges of sex discrimination in college athletics, the defense expert–at great expense to the defendants and the potential confusion of the judge–instituted telephone surveys of student interest using questionable methodology, extensive canvassing of the policies and practices of comparable institutions, and data mining of student application questionnaires when the issue to be addressed by statistical evidence was merely whether the percentage of women in the student body was substantially the same as the percentage of women among the student athletes; fortunately the judge retained his focus on the matter at hand (*Cohen* 1995).

Sometimes charts or graphs are useful, although they can be misleading (Tufte 1997). Sometimes the everyday example that works in an elementary statistics class also works in court. In a recent case, it was clear that lawyers for both sides and the Administrative Judge were baffled by the statisticians' reliance on a hypergeometric model and Fisher's exact test to examine a selection process, especially as in a related case a binomial model of a similar process had been accepted by the court. A simple analogy to drawing aces from a deck of cards with and without replacement brought a dawning of understanding from all involved. Unfortunately, the concept of binomial modeling and the associated Supreme Court-approved measure of disparity in terms of standard deviations (*Hazelwood* 1977) has become firmly entrenched, even where inappropriate.

It is important to remember that statistical–and other–experts must couch their evidence in terms that do not pronounce upon the ultimate legal issue. Thus, statistically significant sex-based disparities in salaries that remain once a number of legitimate factors have been taken into account cannot prove the existence of discrimination; they can only make it appear very unlikely that these dispartities happen by chance. And it must be judged a failure in communication by the statistical experts when courts make such pronouncements as "A fluctuation of two or three standard deviations indicates that the result is *caused* by discriminatory intent rather than chance" (emphasis added; *Ivy* 1986, p. 15).

Lawyers, too, often suffer from a lack of understanding and/or an unwillingness to be instructed; the first task of a statistical expert is often to make certain that the evidence is understood by the attorneys who will be as responsible as the expert for its presentation and reception. Although the outcome of the notorious O.J. Simpson trial probably did not really hinge on the jury's understanding of the DNA evidence, some have blamed what they consider a miscarriage of justice on the jury's innumeracy while

others assert that the prosecution could not explain what they themselves did not understand (Saunders, Meyer, and Wu 1999). In fact, the presentation of the statistical evidence was not a good day for statistics in many ways.

4. Conclusion

Although there is no specific code of ethics binding statisticians, general ethical principles would certainly preclude some statistical presentations that have found their way into courtrooms. Similarly, general principles of good teaching would preclude the complex and muddled presentations that too often appear. It is the role of the statistical expert to convince the finders-of-fact that the evidence can be relied upon. Reliability in a forensic context should include

a) Sensitivity–can the methodology be relied upon to produce results from the quantity and quality of data being examined?

b) Quality control–are factors affecting outcomes understood and controlled for?

c) Discriminating power–can we distinguish between possible outcomes?

plus d) Good old-fashioned honesty–is the presentation such that the judge or jury is convinced that the expert is telling the truth and not tailoring the evidence in order to earn a fee?

Often this last point depends on nothing more than what every good teacher knows: giving the listeners a sense of contact, clarity and confidence combined with absence of evasions and not undermined by superciliousness or arrogance.

REFERENCES

Cohen v. Brown University. (1995). 879 F. Supp. 185 (D.R.I.).

Daubert v. Dow Pharmaceutical. (1993). 509 U.S. 579.

Frye v. United States. (1923). 203 F. 1013.

General Electric v. Joiner. (1997). 66 LW 4036, *reversing* 78 F.3d 524 (11th Cir. 1996).

Gray, M..W. (1993). Can statistics tell us what we do not want to hear? The case of complex salary structures (with discussion). *Statistical Sciences*, **8**, 144-179.

Griffin, C. W. (1999). Dubious "experts." *Washington Post*, January 11, 1999, p. A19.

Hazelwood School District v. United States. (1977). 433 U.S. 299.

Herbert, D. C. and Shelton, L. S. (1996). A pragmatic argument against applying the disparate impact doctrine in age discrimination cases. *South Texas Law Review*, **37**, 625.

Ivy v. Meridian Coca-Cola Bottling Company. (1986). 641 F. Supp. 157 (S.D.Miss.).

Mistretta v. Sandia Corporation. (1977). 1997 WL 17 (D.N.M.), *aff'd sub nom EEOC v. Sandia Corporation*, 639 F.2d 600 (10th Cir. 1980).

McCleskey v. Kemp. (1987). 481 U.S. 279.

Ottaviani v. State University of New York at New Paltz. (1989). 679 F. Supp. 288 (S.D.N.Y. 1988), *aff'd*, 875 F.2d 365 (2d Cir. 1989).

Presseisen v. Swarthmore College. (1978). 442 F. Supp. 593 (E.D.Pa.), *aff'd without opinion*, 582 F.2d 1275 (3d Cir. 1978).

Saunders, S.C., Meyer, N. C., and Wu, D. W. (1999). Compounding evidence from multiple DNA-tests. *Mathematics Magazine*, **72**, 39-43.

Tufte, E. R. (1997). *Visual Explanations: Images and Quantities, Evidence and Narrative*. Graphics Press, Chesire CT.

Wyche v. Marine Midland Bank. (1997). 1997 WL 109564 (S.D.N.Y.).

RÉSUMÉ

If statisticians do not assume responsibility for the clear and convincing presentation of statistical evidence, judicial decision making will suffer. It is important that statisticians understand the role and limitations of statistical evidence and in particular the necessity to convince judges of the validity and appropriateness of their techniques as well of the relevance of statistical evidence. Teaching to the judicial decision makers is an increasingly important contribution that statisticians can make to helping to insure a fair and just legal system.