CRAMMING FOR COURT: TEACHING STATISTICS TO LITIGATORS

Mary W. Gray
Department of Mathematics and Statistics
American University
USA

Experienced litigators pride themselves on being able to do a quick study of any subject, no matter how esoteric. However, self-help frequently does not suffice when the subject is statistics. No matter how well-prepared the expert testimony on their side, the failure of many litigators to be able to understand the statistics of the opposition adequately to cross-examine effectively has doomed many cases. Several examples will be explored, together with guidelines for statisticians who must prep the math-phobic advocate.

For the rational study of the law the black letter man may be the man of the present, but the man of the future is the man of statistics and the master of economics.

Oliver Wendell Holmes
The Path of the Law (1897)

More than a century later, are we in the future contemplated by Holmes? Unlike the situation in many areas of expertise considered by the courts, statistics is a field where all of the players often think they have some understanding of the subject matter, often mistakenly. Courts have frequently accepted what can only be described as pseudoscience and have rejected valid statistical models on spurious grounds. Moreover, rather than the era envisioned by Holmes, the finders of fact all too frequently still operate under Disraeli's principle that there are three kinds of lies: lies, damn lies, and statistics.

If, as is often the case, lawyers are little better prepared to understand statistics than they might have been in the days of Holmes even in the face of an explosion of statistical methodology, just what is the role of the statistician? Certainly ethically and clearly to present the evidence in a case, but the expert cannot ask the next question in a cross-examination challenging invalid alternative interpretations presented by the other side. Hence the expert has an obligation to prepare the lawyer, much as the lawyer has an obligation to prepare the expert for testimony.

The lawyer and the expert must form a partnership, ideally from the beginning so that the statistician is consulted about what data is going to be needed. The statistician and the attorney must understand what questions need to be answered to meet the evidentiary standard in the case at hand and what is required to answer them. Experts cannot testify to a legal conclusion such as whether discrimination has occurred or a conspiracy to fix prices existed but the experts must give, to the extent possible, the litigators the tools to convince the court to come to the correct decision. And, of course, if the litigators actually understands what they are talking about, the advocacy is more likely to be effective.

COMPARISONS OF PROPORTIONS

Fortunately, much statistical evidence is based on relatively simple concepts, nothing more than those facing students in an elementary undergraduate statistics course. Thus it is wise for statisticians to consider their partners a class; in this case the instructors' success, or lack of it, in explaining statistical concepts may affect a million-dollar judgment or someone's life. And, as resistant as attorneys may be, sometimes an appreciation of the principle is best gained by a "modest grappling with the data." (Finkelstein & Levin, 1989, p. ix).

Let us look at a simple example. Many cases hinge on an appropriate analysis of comparative proportions. Consider the case of Cohen v. Brown University, 101 F.3d 155 (1st Cir. 1996), *cert. den.*, 117 S. Ct 1499 (1997), (Gray, 1996). Under the U.S. legislation, Title IX of the Education Act of 1972, discrimination on the basis of sex is prohibited in any education program in an institution receiving federal funds. How to deal with collegiate athletics presents a problem because it is agreed that segregated teams are permitted; that is, for example, there can be a women's swim team and a men's swim team. However what *is* required is that men and women have equivalent opportunities for participation. There is a three-prong test for compliance:

1) Whether intercollegiate level participation opportunities for male and female students are provided in numbers substantially proportionate to their respective enrollments; *or*

2) Where the members of one sex have been and are underrepresented among intercollegiate athletes, whether the institution can show a history and continuing practice of program expansion which is demonstrably responsive to the developing interest and abilities of the members of that sex; *or*

3) Where the members of one sex are underrepresented among intercollegiate athletes, and the institution cannot show a continuing practice of program expansion such as that cited above, whether it can be demonstrated that the interests and abilities of the members of that sex have been fully and effectively accommodated by the present program.

In *Cohen* for the year in question, it was undisputed that 51% of the 5722-person undergraduate student body was women whereas 38% of the 897 students engaged in intercollegiate athletics were women. The defendant university could not show continuing improvement, as the differences in participation rates had not decreased over a period of years. That the interests and abilities of Brown women were not being effectively accommodated was demonstrated by the event that triggered the lawsuit: the cessation of university funding for a previously funded women's team. The legal question presented was: does the 51% to 38% comparison represent substantially proportionate representation? The role of the statistical experts was to assist the court by addressing what "substantially proportionate" means in statistical terms.

Comparisons of proportions have been treated in various ways in litigation. Many employment discrimination cases have involved some particular "test" to be hired or promoted. It could be an actual examination, such as exams given determine eligibility for promotion (*see,* e.g., Connecticut v. Teal, 457 U.S. 440 (1982)) or a hiring criterion such as possession of a secondary school or higher education degree (Griggs v. Duke Power Co, 401 U.S. 424 (1971)). A straightforward characterization of the *Cohen* data would be to point out that there is a 13 point difference in the two figures. However, a simple difference may be interpreted differently depending on the level of the rates. For example, in a case involving jury selection in a death-penalty case the difference between 0% and 5% was deemed sufficient to void a sentence, with Justice Frankfurter declaring: "The mind of justice, not merely its eyes, would have to be blind to attribute such an occasion to mere fortuity" (Avery v. Georgia, 345 U.S. 559, 564 (1953)). On the other hand, in Swain v. Alabama, 380 U.S. 202 (1965), the difference between 26% and 16% was not. It should be noted that the effect of the inexorable zero is observable in many cases and that the factual contexts in which the statistics arose were different.

One can also look at the ratio of the pass (or fail) rates. However, this is subject to manipulation. For example, if 90% of one group pass and 80% of the other group pass, the ratio of the pass rates is 80/90or 89%, but the ratio of the fail rate is 20/10 or 200%. To anyone with a good quantitative sense, that the closer to 100% the pass rate, the more striking is the difference in the pass rate and fail rate ratios may be obvious, but others–including the attorney who needs to recognize the distortion that can result–may need to be instructed. A judge or jury may well feel that a doubled failure rate indicates discrimination but few might believe the same of a 89% comparison of pass rates. In fact, the U.S. Equal Employment Opportunity Commission uses a four-fifths test: a ratio, which is less than 80%, is considered evidence of an adverse impact.

A third descriptive statistic for comparison is the odds ratio. It has the advantage of being invariant; it remains unchanged if both the outcome and the antecedent factor are inverted. Instructors of elementary statistics courses may not be surprised to learn that odds ratio tests find little favor with the courts. Of course, the major responsibility for making clear this or any other statistical concept lies with statistical experts, but if they have failed adequately to instruct their lawyer partners, the partners in turn will be unlikely to be able to play their roles effectively.

Clearly the problem with any of these descriptive comparisons is that no account is taken of the sample size. Sample size–and indeed the concept of sampling itself–presents its own difficulties. Many courts have not realized that most statistical tests incorporate sample size into the calculations and have rejected conclusions based on small samples, no matter how significant the results. Often that some tests are not appropriate for small sample sizes nor certain types of data, has not been recognized, much less the import of the Central Limit Theorem. Other elucidation that an attorney or the court may need is an explanation of the power of a test , the circumstances under which either a binomial or hypergeometric model is appropriate, and the use of one-sided versus two-sided tests. The issue of the power of a test is best illustrated by constructing an example showing the difficulty of detecting a difference with small sample sizes.

Looking at the *Cohen* data for statistical significance, we find that $p < .001$, substantially undermining defendant's claim that the figures met the "substantially proportionate" test, with the "minor" variation being due to chance. Of course, with large sample sizes, much smaller differences in percentages could certainly produce statistical significance, which the court might consider unimportant. The statistician has an obligation to acquaint his partner attorney and ultimately the court with the probability concept involved, leaving, of course the ultimate legal question to be answered by the court. No "bright line" cutoff in terms of any of the tests above has been established in the Title IX cases, nor is there general legal consensus on a threshold level of statistical significance to support a finding of adverse impact. In Castaneda v. Partida, 430 U.S. 482 (1977), the U.S. Supreme Court spoke of "two or three" standard deviations being the rule, without any apparent recognition of the substantial difference between two and three. It should be noted that the probabilities in the two death-penalty cases cited above are .046 in *Avery* and 1×10^{-8} in *Swain.*

Confidence intervals are often an instructive complement or substitute for point estimates in many statistical tests. However, just as in the case of statistics students it can be difficult to get the litigators to understand exactly what they–or the court–are supposed to be confident about. In a recent case involving the geographical distribution of deliveries from an oil pipeline terminus, it seemed impossible to get attorneys to embrace any formulation other than that one could be 95% confident that 80% of the destinations were within 100 miles of the terminus "plus or minus 5%."

OTHER STATISTICAL TESTS

The chi-squared test is often used (and misused) in litigation when comparisons of more than two proportions must be made. Its potential for misleading litigants lies largely in the effect of sample size. As noted above, it is helpful to have in hand an example to show the nature of the problem. For example, if each of the numbers in each cell of a 2 x 2 table is doubled, the value of chi-squared will be doubled, even though the proportions remain the same. The second difficulty with the chi-squared test, or rather a difficulty with those who are not careful about its use, is that it is appropriate *only* for counts, not for measured variables, as has been known to happen.

The standard for the admissibility of expert evidence has traditionally been whether it is generally accepted (e.g., published in a peer-reviewed journal) by the relevant scientific community. Although the threshold for admissibility has been somewhat altered by recent U.S. Supreme Court decisions such as Daubert v. Merrell Dow Pharmaceuticals, 509 U.S. 579 (1993), this is still essentially the guideline. This militates against statistical techniques not widely accepted and may even call into question methodology not easily explained, although that is not its purpose..

More complex techniques are required, or at least preferred, in many situations. Regression is addressed below because of its ubiquity, but ANOVA; the Wilcoxon rank-sum and signed-rank tests; Pearson's correlation and Spearman's rank correlation; Mantel-Haenszel, Fisher and meta-analysis methods for combining data, urn models, and proportional hazards models have all been used. Envision explaining these to someone who may be a quick study but who has never had a course in statistics and who needs to able to spot answers that do not make sense and follow up appropriately!

REGRESSION

One might believe that regression is quite straightforward to understand: we want to see how a dependent variable, such as salary, price, or examination scores, depends upon a collection of independent variables. It has become a widely used tool in cases in such areas as discrimination, antitrust, securities market manipulation, capital punishment, school finance, and environmental damage. However, the shoals upon which regression models can flounder are plentiful.

The selection of explanatory factors requires substantive expertise in the area to be studied. For example, if the issue is whether or not women faculty members at an institution are discriminated against with respect to pay, to set up a regression model requires an understanding of how the particular institution being studied determines pay. Methods can range from a rigid pay scale based strictly upon the years of service to one in which factors like education, productivity and market value, as well as experience, are taken into account. Each of these factors would appear, on the surface, to be sex-neutral and thus a legitimate factor to be used as an independent variable in modeling the salary structure. However, things are rarely that simple. Without going into extensive detail, let us look at the issue more closely. An institution must decide how it values education; it could simply pay a premium of $1000 for any faculty member with a doctorate. But what if a doctorate is not the most appropriate degree for the field of a particular faculty member? Would an MFA be more valuable to an instructor of painting than a doctorate? Should a doctorate in chemistry held by a teacher of mathematics be compensated the same as a doctorate in mathematics? These are, or should be, policy issues determined by the institution. The role of the statistician is not to decide how faculty should be compensated, but rather to take into account in regression model legitimate factors that the institution says it uses plus any other factors that might be used although the institution does not explicitly cite them as contributing to the salary structure.

The rank held by a faculty member almost always is a significant factor in determining faculty salaries; however, clearly if the assignment of rank is determined by the same persons or process that determines salaries, the effect of sex on salaries may appear to be diminished if rank is another of the variables in the regression model; that is rank may be a “tainted” variable. Lawyers can generally understand this, and explain it well, but the fact is that courts have generally discounted any models of faculty salaries in which rank is not a variable, unless there is independent evidence of disparate treatment in the determination of rank. Other possibly tainted variables include rank at hire and administrative experience.

As might be expected, the most frequently used form of regression in litigation is ordinary least squares; in fact, it is so pervasive that a decision that it is inappropriate in a particular situation may require a careful explanation to the litigator. To illustrate how it works it is worthwhile to go through a simple model with a single independent variable an appropriate graphical illustration to help the lawyer understand.

There are some cautions that need to be pointed out; in particular the fact that a few data points with very large residuals can have a major influence on the estimates of coefficients as can explanatory variables with values far removed from the main concentration of data. The partner lawyer may need to verify the accuracy of these data points and/or seek any special considerations that contribute to the large influence. For example, in a recent regression study of faculty salaries it was found that a large positive residual was partly due to an extra year’s salary payout in the year under study as part of an early retirement package.

In attempting to determine whether there is evidence of a disparate impact, the standard method is, of course, to look at the coefficient for sex contained in the model as an explanatory variable. It is a convenient and easily understood differential. Although using logs may produce a slightly better fit, the increase in complexity is generally not worth the effort. The same considerations apply to the use of quadratic terms; because of the diminishing contribution of a year’s experience as the years of experience increase, often the square of experience variables is included. If this is done, the expert needs to make certain that the effect of adding a quadratic factor is clearly understood by the litigator.

Another difficulty with the sex coefficient as a measure of disparate impact is that it actually measures the average difference in salaries assuming that the average effect of all other variables is the same for men and women. That is, were a Ph.D. “worth” $1000 on average for men but only $500 for women, and assuming equal numbers of men and women, the coefficient for possession of a Ph.D. would be only $750 and the underpayment of women would be underestimated. One remedy is to include interaction terms, a concept not intuitively obvious to a lay audience. Another is to produce a men-only regression model and then predict the salaries of women from it, using the average residual as an indicator of disparate impact. In addition to problems potentially caused by loss of power, the use of this model can be considerably more difficult to understand. The pluses and minuses of alternative presentations must, of course, be carefully explained to the litigator.

Although it is easy to explain that the adjusted R^2 is a measure of how much of the variation of the dependent variable is accounted for by the regression model, experts should caution their partner attorney that it may not measure how well the model actually predicts the values of the dependent variable. In fact, although sometimes used as an indicator of the effect of sex in disparate impact cases (*see,* e.g., Wilkins v. University of Houston, 654 F. 2d 388 (5th Cir. 1981), *reh. den.*, 662 F. 2d 1156 (1981), *vacated*, 459 U.S. 809 (1982), *on remand*, 695 F. 2d 134 (1983)), the change in R^2 is not directly related to the percentage shortfall in salaries for women; rather it must be multiplied by a factor that can be very large in order to find the actual shortfall.

In spite of the fact that the statistical significance of a sex coefficient is routinely used as the measure of whether or not the hypothesis of no sex effect can be rejected, if it is not significant this may be due to the lack of power, a high degree of multi-collinearity, or the inclusion of tainted variables rather than to a lack of a sex effect. Moreover, the difference in salaries is real if the whole population is included in the data, not a prediction from a sample. Although it is desirable to have a relatively homogeneous group to study, splitting the population into smaller groups, while it may reduce the standard error, may also reduce the power of the test.

Other complications of regression models need to be clearly understood by attorneys. For example, in the case of faculty salaries, rank, while possibly tainted, may be the only manageable proxy for productivity other than education and experience. Using logistic regression to examine possible disparate impact in assignment of rank has limited efficacy because there are insufficient productivity factors to use as explanatory variables. However, experts and attorneys as well should be aware that when there has been extensive use of variables such as publications and teaching evaluations, their effect is generally slight and sometimes even counterintuitive. Finally, the use of proxies for productivity can in some cases lead to underadjustment bias, another complex concept that may need to be explained to attorneys. In spite of all its limitations in many contexts, regression will no doubt continue to be an effective and widely used tool well worth the instructional time devoted to explaining it.

GUIDELINES

Here are a set of brief guidelines for working as a statistical expert in a litigation context:

1. Make certain the lawyers understand what data are needed *and* why they are needed. They may have to respond to unanticipated questions in oral arguments on motions to compel the production of evidence.

2. Obviously a statistician always has an obligation to select appropriate methodology and to report results accurately and clearly, but keep in mind the necessity to explain first to the lawyer and then to the court what was done, why it was done, and what it means (statistically, not legally).

3. Prepare a script of questions to be asked of yourself and of the other side's expert statistician. Although the guiding principle in litigation is never to ask a question to which the inquisitor does not know the answer, try to anticipate variations that may be needed as a result of unexpected answers.

4. Go over the basic concepts and analysis involved, making clear in particular why one alternative was selected over others. In preparing testimony, statisticians must ask their attorney-partners to explain to them the concept or the significance of the results just explained by the statisticians. If the attorneys do not understand well enough to do so, chances are that the judges or juries will not either.

5. Make certain that the attorney understands the limitations of the analysis so that claims are not made that cannot be substantiated and the statisticians are not pressured to stretch the parameters of what is ethical. In particular, make certain that the attorneys are aware of information adverse to the case.

REFERENCES

Fienberg, S.E. (Ed.). (1989). *The evolving role of statistical assessments as evidence in the courts.* New York, NY: Springer-Verlag.

Finkelstein, M.O., & Levin, B. (1990). *Statistics for lawyers*. New York, NY: Springer-Verlag.

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

Gray, M.W. (1996). The concept of "substantial proportionality" in Title IX athletics cases. *Duke Journal of Gender and Social Policy, 3,*165-188.

Kaye, D. H., & Aickin, M. (1986). *Statistical methods in discrimination litigation*. New York, NY: Marcel Dekker, Inc.