presented by statistics." This is an hypothesis with merit, but one that presupposes the conclusiveness of statistical evidence. Having worked on a number of legal cases involving employment discrimination, I can see several reasons why other aspects of the case may outweigh the statistical evidence.

First, the nature of statistical evidence is largely supportive in legal cases and helps the judge frame a picture of the total evidence. Court opinions hinge on many considerations, including the judge, expertise of the attorneys, testimony of witnesses, documentation of employee policies and numerous legal restrictions, that may override statistical conclusions. I know of at least two cases where fraud and lack of disclosure were dominant concerns and outweighed all other evidence, including the statistical results.

A second reason concerns the adversarial nature of legal proceedings, which often creates tensions and may interfere with statistical conclusions. The court's objective is to get at the truth; but the issues are complex, time is limited and legal matters often restrict the analysis. The statistical conclusions from the opposing sides may be incomplete, exposing two different viewpoints of the data, rather than a more general consensus embracing both. Commenting on truncation bias effects on salary regressions in *Vuyanich v. Republic National Bank* (1981, p. 199), Judge Higginbotham remarked, "Because the controversy here appears to center on an issue on the frontier of econometrics, and there seems, at least to a court unschooled in the intricacies of econometrics, to be genuine conflict between the experts as to the proper approach, we do not decide the issue."

Some attorneys adopt a strategy where the opposing statistical experts simply "cancel each other out." Although this may be appropriate for the legal case, it hardly advances our scientific knowledge about the employment process. But it is a reality of litigation and one that confronts statistical experts for both plaintiffs and defendants.

Finally, there are formidable difficulties in statistical assessments of discrimination from employment data. We do not have a well-developed economic model of the employment process that yields definitive measures of discriminatory departures. Bloom and Killingsworth (1982), Ehrenberg and Smith (1985) and Cain (1986) document the current level of knowledge and evolving economic theory. The intent of Title VII has been to prevent unequal treatment of individuals but also not to infringe on differential employment decisions that arise from job relatedness or "business necessity" (Fiss, 1979). What constitutes "business necessity" is not a simple matter and varies considerably across different organizations. The most recent legislation for the 1991 Civil Rights Act affirmed the importance of "business necessity," while deliberately leaving it vaguely defined and for individual cases to decide.

There is a great need for a well-developed model of the employment process and accurate measures of relevant income, job qualifications and market-related factors. Although we have made substantial progress, the appropriate models and data bases are still evolving. The work to improve inadequate data sources is hardly glamorous and requires painstaking effort. Nonetheless, only with accurate data bases and more complete structural models of the employment process can we hope to achieve better statistical measures of discrimination. In the meantime, we must strive to maintain scientific objectivity and a balanced viewpoint, summarizing the data as best we can and acknowledging areas of uncertainty, to promote informed policy decisions.

ACKNOWLEDGMENTS

This work was partially supported by the School of Business Administration Research Fund, University of Southern California. The author would like to thank Lou Gordon and Ingram Olkin for helpful conversations and suggestions.

Comment

Joseph L. Gastwirth

Because of the author's dual qualifications as a lawyer and statistician, it was interesting to observe that two themes of the article were the view that statistical studies are used to bolster decisions that the policymaker or judge was leaning to anyway and the reluctance of judges to rely solely on statistics. After first commenting on these and other general issues raised by Professor Gray, I will then discuss the regression analyses used in some of the cases cited. As I previously participated in the discussion of Professor

Joseph L. Gastwirth is Professor, Statistics Department, George Washington University, 2201 G Street, N.W., Washington, D.C. 20052.

Dempster's article on errors in variables, less attention is devoted to that area. Finally, a few general suggestions for statisticians involved in the legal process are offered. The implications of the 1991 Civil Rights Act for statistical evidence will be given attention.

1. GENERAL COMMENTS

While the author's skepticism about policymakers and the judiciary is justified in many ways, I would like to demur on her comments asserting that the burden the *Ward's Cove* case put on plaintiffs was to show that *no factor other than race* could explain a disparity in pass or hire rates and the statement that the *Allen v. Seidman* and *Green v. USX* courts may have been guided by different ideological considerations. The *Ward's Cove* decision made it harder for plaintiffs to prevail in disparate-impact cases and led to a noticeable decline in the fraction of civil rights cases plaintiffs won (unpublished data from Dr. J. Cecil of the Federal Judicial Center indicating a decline of 10% is quoted in Gastwirth, 1992b). The majority opinion did not state that plaintiffs have the burden of showing that no factor other than race could be responsible for an observed disparity as suggested in the opening section of the article. The Court required plaintiffs to make a focused comparison of similarly qualified individuals in order to make a *prima facie* case. The *Ward's Cove* decision also made it easier for defendants to justify an employment practice having a disparate impact but did subject the explanation to a reasoned review. A number of post-*Ward's Cove* cases are discussed in Gastwirth (1992b), and we assess the effect of the new Civil Rights Act on statistical testimony later in section 3.

The *Allen* court was composed of three judges who had been appointed by President Reagan and was written by Judge R. A. Posner, a leading exponent of economic analysis of the law as well as jurisprudence in general. The opinion in *Green* was written by Judge A. L. Higginbotham, a Carter appointee who has written extensively on the situation of blacks in South Africa. It is not apparent to me that the two panels have a common ideological view, which also is contrary to the ideology of the majority (five Republican and one Democratic appointee) in *Ward's Cove*. More importantly the statistics in both cases showed that whites had twice the success rates of blacks. An evaluation of the data for the possible effect of an omitted variable (OV) using Cornfield's approach (Greenhouse, 1982) indicates that the OV would have had to double an applicant's probability of passing the test in *Allen* or of being hired in *Green* and would have to be twice as prevalent among whites than blacks (see Gastwirth, 1992a, b, for details). The *Allen* opinion observed that all applicants had served for over five years at the FDIC, a year or more at the previous grade level (GS-9) and received a recommendation from the supervisor in order to take the exam. Thus, the black and white applicant pools appeared to be homogeneous. The opinion specifically noted that the defendant had the opportunity to reanalyze the data, using a regression analysis (presumably logistic) incorporating the variable or factors the plaintiff omitted but did not choose to do so. Indeed, the court remarked that for all it knew a regression might yield a larger estimate of the negative effect of being black had than the simple comparison. I believe these cases are consistent with the *Bazemore* decision which not only approved of the regression methodology, as stated by Gray, but also said that plaintiffs do not have to incorporate *all* measurable variables although the major ones should be used.

Professor Gray points out the difficulty plaintiffs have encountered in winning tenure in Title VII cases. Indeed, one case *Ford v. Nicks* (1987) was overturned on appeal in 1989. On the other hand, plaintiffs have been more successful in two recent cases. *Benunn v. Rutgers University* [941 F. 2d 154 (3d Cir. 1991)] and *Jew v. Iowa State Medical School* [57 FEP Cases (S.D. Ia. 1990)], where the plaintiffs, an Hispanic male and Chinese female respectively, successfully challenged their departments' failure to promote them to full professor. Since the Supreme Court in *Watson v. Fort Worth Bank and Trust* allowed subjective practices to be scrutinized under the disparate-impact approach, plaintiffs may have more success in the future. The author's skepticism concerning the prior beliefs of judges and policymakers is supported by a dissent in *Benunn* by a usually pro-plaintiff judge who taught at a university before being appointed by President Carter. Judge Sloviter's dissent states, "the majority's decision may be read to thrust the federal courts . . . into the subjective area of academic tenure and promotion to an unwarranted and unprecedented degree." I urge interested readers to examine both the district and appellate opinions, which describe the comparison of the plaintiff with another recently promoted faculty member. The second faculty member was a "good teacher" whose distinctly lower research output received a better rating by the department than the plaintiff's. The appellate majority noted that while Rutgers articulated five relevant criteria, the promotion process was so undermined by inconsistency in their application that the district court's finding of discrimination was upheld. Since the Supreme Court, on January 21, 1992, declined to consider Rutgers' request that it consider the case, I am unsure about the full impact the case will have.

The *Benunn* case indicates that Professor Gray's discussion of which factors should be incorporated in statistical comparisons and the relevant time frames are important issues for the statistical community to consider. While the *Allen* and *Green* decisions were well reasoned, some courts have allowed defendants to use factors like "interest" to explain a statistically significant and meaningful disparity. Perhaps the best known case where this occurred is *EEOC v. Sears* [839 F.2d 302 (7th Cir. 1988)]. Although the firm asserted that women were not interested in certain sales positions, shortly after the charge was filed, the fraction of females hired into these jobs rose dramatically (see Gastwirth, 1992b, for details). This "A Funny Thing Happened on the Way to the Courtroom" phenomenon occurs all too often and is a major reason plaintiffs have not prevailed in several cases (Gastwirth, 1988). It should be noted that the opinion in *Green* used a sharp increase in black hires shortly after the charge to support an inference of discrimination, as it indicated the availability of qualified blacks in the labor market. On the other hand, the recent decision in *EEOC v. Chicago Miniature Lamp Co.* (57 FEP Cases 408 1991) again accepted the defendant's assertion that blacks were not interested in the low-paying job without supporting data. The majority opinion in *Chicago Lamp* raised serious questions about the labor market constructed by the plaintiffs and could have reversed the lower court or remanded the case for further fact finding on the issue without accepting the "lack of interest" assertion. It is interesting that the case was decided in the same Circuit as *Allen* and *U.S. v. Rockford Memorial Corp.* [898 F.2d 1278 (7th Cir. 1990)], an antitrust case concerning the monopolistic effect of a proposed merger of hospitals. The statistical issue concerned the construction of the market area served by the hospitals. Analogous to the local labor markets used in civil rights cases, the Justice Department created a market area assuming that individuals prefer hospitals closer to where they live. When the defense asserted that a much larger area was appropriate, Judge Posner pointed out that it is easy to take "pot shots" at a proposed market area, but the *criticisms need to be substantial.* The opinion characterized the defendant's claim that an *unweighted* area including all ten counties from which the hospital drew any patients was "ridiculous" as 87% of the patients resided in three counties.

As noted by Professor Gray, similar issues arise in determining which variables are proper measures of productivity and need to be considered in a wage regression, and which variables may simply reflect discriminatory practices of the employer. The discussion of administrative experience, especially of *prior* rather than *current* experience, as a legitimate factor is very important and may apply to service on university committees as well. I am unaware of any university that has a systematic procedure designed to ensure that all faculty members, especially the newer ones, are given an opportunity to explore their potential for administrative tasks. Furthermore, is it appropriate to pay a *former* administrator, who has returned to the classroom more than an otherwise comparable professor for the same work?

Professor Gray's treatment of the evaluation of faculty research and scholarly and creative activity is quite instructive. Of all the criteria used in the promotion process, this should be a fairly objective one. After all a department could group journals into three or four categories according to their prestige and impact on the profession and distribute the list to all faculty members. As a member of the Personnel Committee of my own Arts and Sciences College, I have found that departments are very reluctant to describe the approximate standing of journals. Some need to be pushed to distinguish between papers appearing in conference proceedings and those published in peer-reviewed journals. While journal rankings are far from perfect indicators, they surely are superior to letting the chair of a department evaluate a faculty member's research without reasonably specific guidelines. A similar potential for favoritism or unconscious bias exists in the process of selecting outside evaluators.

2. STATISTICAL ANALYSIS IN ACTUAL TITLE VII CASES

Professor Gray's discussion of the problems inherent in developing regression analyses appropriate for EEO cases is a helpful introduction to the subject. As a substantial literature [see Dempster (1988) and Finkelstein and Levin (1990) for references] exists in this area, I will focus on two cases cited by Professor Gray.

In *Ottaviani* and *Denny*, plaintiffs used the male-only or Peters-Belson approach to predict female salaries while the "dummy variable" was used by the defendants. As noted elsewhere (Gastwirth, 1989) the two approaches only estimate the *same* parameter, the difference in intercepts when regression planes agree, that is, the assumptions for the validity of the "dummy variable" approach are satisfied. Otherwise, the average difference ($\overline{D}_1$) between the actual female salaries (Y_i) and their predicted values derived from the male equation estimates

$$\delta = \overline{X}_1' (\beta_1 - \beta_2), \tag{1}$$

where X_1 is the mean of the covariate values of the females and β_1 and β_2 are the coefficients of the female and male regressions, respectively. This parameter, δ, reduces to $(\alpha_1 - \alpha_2)$, the difference in intercepts when the coefficients of the covariates are the same. Of course, $\overline{D}_1$ being the average difference in salaries between what females actually receive and what they would be paid had they been compensated according to the male formula is a very meaningful statistic in this context, whether or not the planes are parallel. When the regressions are *not* parallel, the coefficient, γ, of the "dummy variable" estimates the *difference* between female salaries and their predicted values, not at their average covariate value but at a weighted average of covariate values of the males and females. Moreover, these weights depend on the variance–covariance matrix of the covariates. Thus, γ is not an intuitive parameter under these circumstances. The bottom line is that, unless the regression planes are parallel, the two procedures estimate *different*, although related, parameters. Since this point is rarely mentioned by the parties, it is surprising that courts get confused? Judges then turn to an assessment of the variables incorporated in the models and an intuitive evaluation of which model is most sensible.

In *Denny*, the plaintiffs developed the male-only model by using the following covariates: departmental groups, seniority at the college, years of prior experience (both tenure track and other), current rank and the number of years since their terminal degree was received. When the female salaries were predicted from the male equation the yearly shortfall ranged from about \$500–\$700 in the early years (1974–1981) to \$1,300–\$2,000 in 1982–1984. The defendant's expert criticized the model asserting that it suffered from the technical difficulties of "multicollinearity," "nonrandomness of the residuals" and "heteroscedasticity." The opinion notes that the expert had not made a study, that is, performed a statistical test of the non-randomness of the residuals, he just assumed it from prior experience. Judge Freedman then discusses the effect of multicollinearity crediting the plaintiffs' expert's explanation that multicollinearity affects the precision of the estimates of the coefficient, thereby *overestimating* the standard error, which would increase the p-value of the test of significance. Hence, he did not credit the criticism.

Of greater relevance is the judge's assessment of the model, in particular, the way the departmental groups were created. The opinion accepted the related subject matter approach of the plaintiffs but expressed concern about the possible effect of combining one department with a very high anticipated wage (psychology) with one of the lowest (sociology) in the Social Science group. The defendants questioned the plaintiffs' expert for not including distinguished service awards (DSAs) in the equation; however, when they were included (in a trial after cross-examination, the party putting the witness on gets another chance, called redirect exam, to rebut questions raised in the cross-exam), the results were not significantly changed.

The defense also introduced a regression analysis, which showed a statistically significant "sex coefficient" in 5 of the 11 years. Although the opinion does not list the covariates used, it notes the multicollinearity problem and the tendency for this to inflate p-values, thereby making truly significant results nonsignificant. The defendant attempted to explain the few years with the statistically significant wage disparities by the rapid expansion of two departments, computer science and business between 1980 and 1984. The court observed that this could not explain the plaintiffs' results in the earlier period and that the defendant's model included a dummy variable indicating membership in the computer science department, so they should not double-count it. Although statisticians could question whether a single-dummy variable completely captures the special effect of such a factor (e.g., being in computer science and possessing a Ph.D. or having a certain type of prior experience may justify an additional increment), we can see that Judge Freeman is seriously considering the potential impact of any suggested "flaw," quite consistently with Judge Posner's opinion in *Allen* and Judge P. E. Higginbotham's classic *Vuyanich* opinion. Moreover, the judge is not supposed to develop the statistical analysis. That is the duty of the parties.

The regressions presented in *Ottaviani* are discussed by Professor Gray, who was the plaintiffs' expert. The opinion noted that the demonstration in Table 3 of the variability of the year of hire coefficient was a cogent criticism of the defendants' model. However, it also noted that the defendants' logistic regression showed that women had *not* suffered discrimination in promotion. Hence, *rank* was an appropriate covariate. It would be helpful for Professor Gray to report the full regression as well as the separate regressions. The reader could then assess whether the estimated coefficients make economic sense. The dollar value attached to prior experience or to being a chairperson could also be examined. Professor Gray did include a dummy variable for department chair and found that it reduced the sex coefficient to nonsignificance in only two years. Had time and resources been available, perhaps the plaintiffs could have incorporated both administrative factors and deleted the year-of-hire variable. Even if this model yielded a nonsignificant coefficient in some of the years, it may well have been significant in a majority of them and especially in the early years. Under *Bazemore*, they would deserve monetary relief for those years. Had this proposed regression indicated that the two administrative factors reduced the coefficient for sex to nonsignificance in virtually all years, this would shift the focus of inquiry to the fairness of the process used in making these appointments. Presumably an analysis comparing the fraction (zero in the case of non-chair administrators) to the female fraction of either actual applicants (if the data are available) or *eligible* employees as was done in *Hogan v. Pierce* could be carried out. If there were at least 20 such positions then, and females formed at least 20% of the eligible faculty, the fact that *no* woman ever received an administrative appointment might well be statistically significant. Without the number of individuals having such experience and the female fraction of the eligible pool, it is difficult to know how much weight to give the *zero*. I know of no case in which a statistically significant *zero*, obtained by a careful comparison along the lines indicated, was not ultimately recognized as strong evidence of discrimination by the courts.

When the linear model is appropriate, the two-equation approach suggested by Gray and Scott (1980) not only helps one to study whether the two regressions are the same (Finkelstein, 1980), it also enables us to explore the validity of the assumption of equal error variances. Unfortunately, these procedures still can be affected by errors-in-variables (EV) problems (Peterson, 1986), and most existing procedures for accounting for EV problems require some knowledge about the measurement error covariance matrix, Ω, (Schafer, 1987). Thus, further exploration of the applicability of these methods to actual data should prove quite useful. One might be able to demonstrate that a "reasonable approximation" to Ω suffices for many purposes. There is another technical question when the differences $\overline{D}_1$ ($\overline{D}_2$) obtained from comparing the female (male) salaries to the predictions from the male (female) equations. As the same observations are used to fit both equations, $\overline{D}_1$ and $\overline{D}_2$ are statistically dependent. The correct variance of $\overline{D}_1 - \overline{D}_2$ would help in assessing the consistency of these estimated differentials.

Another problem with the "dummy variable" total-population approach is that the estimated coefficient for "sex" not only reflects the effect of "sex" but the residual effect of any excluded variables, not fully accounted for by the covariates. Indeed, the conditions (Pratt and Schlaifer, 1984) for the interpretation of this coefficient as the causal effect of sex are usually not satisfied in observational studies. A recent paper (Swamy and Tavlas, 1992) reviews and extends the ideas of Pratt and Schlaifer and applies them to analyze money demand. In our application "sex" should be a factor rather than a concomitant, but, according to Pratt and Schlaifer, if a variable serves as a proxy for an excluded one, it should be a concomitant. These considerations imply that the appropriateness of a statistical model used in a discrimination case depends on whether or not the *appropriate* job-related variables are included in the model. Of course, failing to include an important covariate also diminishes the relevance of the estimate $\overline{D}_1$ in the Peters-Belson approach.

3. THE POTENTIAL EFFECT OF THE CIVIL RIGHTS ACT OF 1991 ON STATISTICAL EVIDENCE

Although Professor Gray submitted the article before the new act became law, it is important to assess the effect it may have on future cases and on the role of statistical evidence in civil rights litigation. In the *Ward's Cove* (1989) case, the Supreme Court changed the prevailing standards of proof for disparate-impact cases to bring them closer to disparate-treatment cases. (The plaintiff always had the burden of proof in disparate-treatment cases; once they demonstrated unequal treatment the defendant only needed to produce evidence sufficient to show that a legitimate factual issue remained. The plaintiff then could show the defendant's justification was a pretext.)

Disparate-impact cases concern the effect of a single or related set of practices on a protected group. The standards were established in the *Griggs v. Duke Power Co.* (1971) case. In brief, the case dealt with the effect on blacks of the firm's requirement that applicants for blue-collar jobs had a high school diploma or passed an intelligence test. The data are given in Gastwirth (1988, p. 261). Because a much lower fraction of blacks met the requirement than whites, the Court required the defendant to demonstrate that it was a valid predictor of job performance, that is, once the plaintiffs showed an employment practice had a disparate impact, *Griggs* placed the burden of proof of the "*business necessity*" of the practice on the employer. If the employer met that burden, the plaintiff had an opportunity to show that an alternative practice, having less of an impact on minorities, could serve the employer's needs. The *Ward's Cove* case, described in Gastwirth (1992a, b), required plaintiffs to demonstrate the disparate impact of a *single practice* by a fairly precise comparison of similarly qualified minority and majority applicants. If they established a *prima facie* case, the *Ward's Cove* decision placed a burden of producing evidence that the practice at issue served a *legitimate business* purpose on the defendant's part. Thus, the decision not only made it more difficult for plaintiffs to establish that an employment practice had a disparate impact, it made it easier for defendants to justify such a practice. The 1991 Act preserved the requirement that plaintiffs make a well-focused comparison in order to establish the disparate effect of a practice but restored the *Griggs* "business necessity" standard and I believe restored the placement of a serious burden of production on the defendant on the issue of the validity of a practice having a disparate impact. Since the lower courts have split as to whether the 1991 Act applies retroactively, a majority of the courts that have considered the issue have said it does not. When reading recently issued decisions in disparate-impact cases one needs to ascertain which legal standards are applicable.

These changes imply a greater need for reliable data concerning the qualifications of applicants in order for plaintiffs to assess the impact of a practice and for accurate job-validation studies of practices which exclude a disproportionate fraction of minorities from job opportunities. Elsewhere (Gastwirth, 1992b), I indicate that the statistical community will need to put pressure on the government to provide more detailed census data for smaller areas and subareas of metropolitan statistical areas than is now the case and that the EEOC should require firms to preserve employment data for several years.

Perhaps of more long-term importance is the need for statisticians working in this area to refocus their attention from a concentration on regression methods based on the linear model to more flexible nonparametric regression techniques and the analysis of 0-1 data, that is, the comparison of hire and promotion rates. Although methods such as logistic regression and stratification (the Cochran-Mantel-Haenszel procedure) have been accepted by courts, relatively little research has been carried out on combining logistic regression with stratification and the effect of omitted or mismeasured covariates on these procedures. In particular, strata formed using the educational background of applicants or employees is subject to more mismeasurement than strata formed by date of hire. Analogs of Cornfield's result and the related results of Rosenbaum and Rubin (1983) and Rosenbaum and Krieger (1990) are needed here. It would be helpful to courts if we could provide an approximate classification of the potential effect of various flaws in data. In this way judges would be on notice to question the reliability of certain types of data or analyses more carefully.

The new act may have an unintended side effect on statisticians testifying in these cases; namely, the act *allows* plaintiffs to ask for *jury trials*. This means that each side will be offering evidence to persuade a jury, which roughly can be regarded as a sample of the population in the local area, but will need to create a record strong enough to be upheld on appeal. Of course, appellate judges cannot overturn a jury verdict solely on the basis that they disagree with it. They need to find that there was not a rational basis for the decision. Nonetheless, both sides will be attempting to solidify their presentations for two different audiences.

I believe that methods based on matching similarly qualified minority and majority applicants or forming appropriate strata will be more understandable to juries than regression models with many variables. Nonparametric regression approaches such as that of Bhattacharya (1989), based on forming a band in the covariate space around each minority member and comparing the minority and majority members in the band with the Wilcoxon or difference in means and then pooling the results, will be easier to explain. Actually Bhattacharya's method also uses groups formed by a band about each majority member and is extendable to using the Scott-Gray idea of looking at the two comparisons separately to ensure a logical consistency. Notice that the problem of widely different experience levels or values of another covariate in the two groups originally noted by the late Professor DeGroot (McCabe, 1980) and properly stressed by Professor Gray could cause bandwidth matching to break down in an extreme situation. In the long run it is better for our profession to say that the data do not permit a sound comparison of similarly qualified individuals rather than build an analysis based on an unverifiable assumption, such as a linear relation between salary and job tenure that has to be extrapolated far from the range of the data in the given sample. Of course, other approaches to nonparametric regression should be considered, and it would be of interest to create and fit the model with the majority group data and then use the Peters-Belson approach to predict the minority salary. Indeed, when the linear model holds, juries are more likely to understand the meaning of the parameter, δ, than the "dummy" variable. An advantage of Bhattacharya's technique is that it extends to the 0-1 case, as Bhattacharya and Gastwirth (1989) obtained an analog to a common odds ratio after adjustment for the covariates used in the matching process. Of course, matching methods are not a panacea as they are also subject to EV and OV problems. The OV problems should not be so important that the employer should obtain data on the major productivity characteristics.

Another potential but more problematic area is the introduction of Bayesian methods. In discrimination cases, unlike in criminal cases where evidence of prior criminal activity is inadmissible for the purpose of showing that the accused has a propensity for crime to avoid bias—which means that relevant knowledge upon which a sound prior could be based is excluded—almost all relevant facts can be submitted. It is possible for jurors to base their prior on the nonstatistical evidence, the demeanor and testimony of the plaintiff and employer's staff, the work history of the plaintiff and comparable employees and the EEO record of the defendant for several years before the charge as well as the time sequence of events in the case. Kadane (1990) shows the applicability of the Bayesian approach for the combination of 2 × 2 tables. Although the articles by Saks and Kidd (1980), Cecil, Hans and Wiggins (1990) and Kaye and Koehler (1991) indicate that juries have difficulty with probability evidence in general and Bayesian methods in particular, ways to utilize the Bayesian paradigm to combine various sources of relevant information deserve more attention.

4. SPECIAL PROBLEMS ARISING IN THE PRESENTATION OF STATISTICAL TESTIMONY

Gray's article reminds us of the literature (Gibbons, 1973; Meier, 1986; Fienberg, 1989) concerning special ethical problems occurring in the adversary process. Rereading the *Ottaviani* decisions suggested to me that statisticians participating in the legal process as experts or consultants could use some guidance from Professor Gray in order to avoid being caught up in the adversary nature of the proceedings and pitfalls created by the legal rules of procedure and evidence.

In *Ottaviani*, the plaintiffs' expert first asserted that a data set was too small to analyze. I believe the defendant's expert agreed. Subsequently, the plaintiffs desired to apply a formal statistical test to the data, but the court did not allow them to. Presumably, procedural rules designed to ensure fairness to both parties justify the court's decision. A similar situation arose in another case when at a pre-trial deposition an expert asserted that a 2 × 2 table should be analyzed by the chi-square test. Because of the small sample size, at trial the expert desired to use Fisher's exact test, as the computer output for the chi-square included a warning that the expected cell count was less than five in some cells so the conditions for the validity of the chi-square approximation were not satisfied. Again the court did not allow this testimony as the opposing side could not be prepared for a proper cross-exam. While new computer programs such as STATXACT may alleviate the small sample-size problem, as the data set can readily be analyzed, new approaches often occur to us after we make our first analysis. How can statisticians, especially at pre-trial depositions, appear knowledgeable and yet leave the door open for alternative analyses to be given later at trial? The problem with small samples is their low power to detect meaningful differences. Unfortunately, courts have often failed to appreciate this. With STATXACT and other programs (Goldstein, 1989) hopefully we will be more persuasive in future cases.

The ethical constraints on lawyers differ from those of academia, and experts face a number of unusual problems (Fienberg and Straf, 1991). Should one carry out an analysis that will likely not be in the best interest of the client? Should one do something that the lawyer should not do because it violates their ethical canons? A problem I have faced is the existence of other data sets that the lawyer did not tell me about. When analyses of the new data are submitted by the opposing party we do not have time properly to assess the comparative reliability and relevance to the issue at hand of the two data sets. The lawyer who has put you on the stand desires you to criticize the "new" data set, for example, to point out that some data are missing, some applicants are counted twice and so on. Statistical experts might well wish to avoid commenting without studying the data for a while, and it is tempting to assert that one should avoid any testimony. However, some of the flaws just cited may apply to the new data set. Is it fair to the court not to point them out? Is there a way to obtain a reasonable amount of time to carry out an assessment of the data? Remember, the lawyer who hired you did not tell you about it, so assume it will not help the party that hired you. I am unaware of any way prospective experts can assure that they will be given all the data relevant to the issue they are asked to study before the trial. I hope Professor Gray might offer some suggestions for avoiding these problems.

Comment

Harry V. Roberts

INTRODUCTION: WHAT IS THE RIGHT QUESTION?

When I first looked at the title, "Can Statistics Tell Us What We Do Not Want to Hear?" my reaction was, "Only with great difficulty." Professor Gray almost immediately echoed my reaction by saying, "It often appears that the most, indeed perhaps the only, effective role of statistics is to bolster decisions policymakers were prepared to take on other grounds." She added, "A corollary to the assertion that statistics are believed only when they conform to how one wants the world to look is the theory that the more closely statistics challenge one's own interest, the less likely they are to be relied upon."

The specific testing ground is the area of employment discrimination, with alleged salary discrimination against female faculty members as the principal illustration. Professor Gray provides a lucid overview of the problems in using statistics – regression analysis in particular – to illuminate the legal question of whether or not discrimination against females, minorities, or other protected groups has occurred. Her description of the evolving legal background, groundrules and guidelines for the use of statistics in discrimination cases is most helpful. The difficulties and seemingly erratic variations in the response of courts to statistical argumentation are skillfully and accurately depicted.

Harry V. Roberts is Professor, Graduate School of Business, University of Chicago, 1101 East 58th Street, Chicago, Illinois 60637.