Statistical Science
1993, Vol. 8, No. 2, 144-179

Can Statistics Tell Us What We Do Not Want to Hear? The Case of Complex Salary Structures

Mary W. Gray

Abstract. 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. The secondary effects of smoking, the sex differential in SAT scores, the census undercount controversy, the validity of DNA evidence and the evidence of the relation of the race of the victim to the imposition of the death penalty all provide examples of the intertwining of political and statistical considerations. Although the court in *Alabama v. United States* (1962) attributed to statistics the power to speak, if courts do not want to hear what the statistics are saying they just do not listen.

Moreover, it sometimes appears that statistics are less likely to be relied upon if they challenge one's own interest. Confronted at their own institutions with an analysis of faculty salaries showing evidence of discrimination, often the very faculty whose stock-in-trade is to persuade others of the efficacy of statistics refuse to believe what is presented to them.

Courts share this skepticism; this paper examines the courts' treatment of the statistical issues involved in studying possible discrimination in faculty salaries.

Key words and phrases: Discrimination, faculty salaries, multiple regression, judicial reception of statistics.

1. INTRODUCTION

That television is mercifully free of cigarette commercials is the result of a policy decision based on statistics showing a strong correlation between cigarette smoking and lung cancer. More recently, knowing that even very accurate tests will identify a large number of false positives when a disease is rare has so far helped to discourage routine HIV testing of populations at low risk for AIDS (Gastwirth, 1987). Statistics-based policy decisions routinely are being made in the establishment of workplace health standards, in the Food and Drug Administration drug approval process and in market investment strategies.

Controversy continues over the soundness of the statistical methodology used to measure the secondary effects of cigarette smoking (Blot and Fraumeni, 1986; Mantel, 1990). However, the question may be largely academic because policymakers appear to have decided on political grounds to move toward a smoke-free environment. 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. It has been said that statistics play the same role for policymakers that a lamp pole does for a drunk: it provides support rather than illumination. The evidence that allowed the Surgeon General to make the determination we find on cigarette packs probably would not have been considered conclusive were the tobacco industry as important to the nation's economy as it once was.

On each year's Scholastic Aptitude Test (SAT) college-bound high school senior boys outscore girls by more than 50 points. With the hundreds of thousands who take the test, the probability of such a difference occurring by chance is less than one in a billion. Studies have shown that the SATs by themselves are not a good predictor of success in college (Clark and Grandy, 1984; College Board, 1988; National Commission on Testing and Public Policy, 1990). These two pieces of evidence caused a New York court to order the state to stop using the SATs as the sole means of awarding

Mary W. Gray is Professor, Department of Mathematics and Statistics, College of Arts and Science, The American University, 4400 Massachusetts Avenue, N.W., Washington, D.C. 20016-8050. She is also a member of the District of Columbia Bar.

state scholarships to high school seniors, asserting that the tests are discriminatory (*Sharif v. New York State Department of Education,* 1989). Although the judge no doubt was concerned that worthy young women not be deprived of a chance for higher education, the decision could be seen more as a reaction to the tyranny of standardized tests than as a tribute to the statistical evidence.

Central to our system of governance is apportionment, determined by a constitutionally mandated decennial census. The Census Bureau and many statisticians favored a statistical adjustment of the 1990 census undercount (Wolter, 1991; Wolter and Causey, 1991). Although there were statisticians to be found on the other side (Freedman, 1991), when Secretary of Commerce Robert Mosbacher made the decision against adjustment, it was widely perceived as being politically motivated (Barringer, 1991).

The Supreme Court refused to declare the death penalty unconstitutional in the face of statistical evidence that its use was related to the race of the victim (*McCleskey v. Kemp,* 1987; see also Baldus, 1990). The Court held that the statistics were not relevant to whether the jury in the particular case before it was impermissibly influenced by considerations of race. Many believe that no evidence of bias would be effective against the pro–death penalty stance of the majority of the current Supreme Court. Although the court in *Alabama v. United States* (1962) attributed to statistics the power to speak, if courts do not want to hear what the statistics are saying, they just do not listen.

Although much of the discourse about the DNA evidence centers on whether the methodology being employed is appropriate, lurking underneath is the more fundamental question of the role of statistical evidence (Cohen, 1990; *Commonwealth v. Curnin,* 1991; Hoeffel, 1990; Kaye, 1989; Neufeld and Colman, 1990; *People v. Castro,* 1989; Tande, 1989). Doubts about whether guilt or paternity should be determined solely by statistical evidence overshadow doubts about the statistics themselves. Recent allegations of FBI and prosecutorial efforts to suppress challenges to DNA evidence have complicated the issue (Chakraborty and Kidd, 1991; Lewontin and Hartl, 1991; Roberts, 1991).

In the employment discrimination arena we are suffering a backlash against the use of statistics. In vetoing the Civil Rights Act of 1990, President Bush claimed that the bill so unfairly exalted the role of statistical evidence of discrimination as to induce employers to impose hiring quotas as their only means of defense. However, the bill he signed the following year differed little in this respect.

In one of the decisions whose reversal Congress sought in the vetoed civil rights legislation, the Supreme Court apparently put the burden on plaintiffs to show that no factor other than race could be responsible for an observed disparity in hiring rates between whites and minorities (*Ward's Cove Packing Co. v. Atonio,* 1989). Subsequently lower courts, perhaps guided by different ideological considerations, have backed away from this standard (*Allen v. Seidman,* 1989, p. 381; *Green v. USX Corporation,* 1990, p. 805). Imagine where the Surgeon General would be if she had to prove that no conceivable other factor could contribute to the differential mortality rates between smokers and nonsmokers.

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. Confronted at their own institutions with an analysis of faculty salaries showing evidence of discrimination, often the very faculty whose stock-in-trade is to persuade others of the efficacy of statistics refuse to believe what is presented to them.

BACKGROUND

Title VII of the Civil Rights Act of 1964 outlawed employment discrimination on the basis of race, sex, national origin or religion (with certain exceptions). In *Griggs v. Duke Power Co.* (1971), a decision essentially reversed by *Ward's Cove,* the Supreme Court identified and outlawed what has come to be known as disparate-impact discrimination, that is, the use of a nominally neutral employment standard which adversely affects a protected group and which is not required by business necessity. The criterion in question in *Griggs* was the requirement of a high-school diploma for a job as a power line repair person; a comparison of the percentage of the adult male white and African-American populations who were high school graduates provided evidence of the requirement's disparate impact on African-Americans. Plaintiffs also showed that among whites employed prior to the institution of the degree requirement a high-school education was unrelated to job performance.

Following *Griggs,* statistical evidence came to play an important role in employment discrimination cases, both those based on a disparate-impact theory and the more traditional discrimination cases where disparate treatment of similarly qualified individuals was alleged. Typically a comparison would be made of the percentage of minorities in an availability pool to the percentage actually hired or between the mean salary of minorities and the mean salary of whites or between the salaries of men and women; in addition to Title VII prohibitions, the federal Equal Pay Act forbids gender-based discrimination in compensation. The provisions vary somewhat from those of Title VII. In making these comparisons, the Equal Employment Opportunity Commission (EEOC), the agency charged

with the administration of Title VII, propounded the so-called four-fifths rule: a selection rate for minorities less than four-fifths of the highest selection rate is taken to constitute evidence of discrimination. Courts were never very enamored of this rule, and statisticians soon convinced them to substitute a measure of statistical significance. The Supreme Court blessed "two or three standard deviations" (*Hazelwood School District v. United States,* 1977) as an appropriate guideline, with no apparent understanding of the enormous range this offhand remark implied. About the only remnant of the four-fifths rule is the courts' caution with small samples, which appears to be based on a failure to realize that unlike in the case of the blanket four-fifths rule, a calculation of statistical significance already accounts for sample size (Gastwirth, 1988).

Notwithstanding the Supreme Court's enunciated standard deviation rule, lower courts have applied varying standards or failed to endorse a particular cutoff point, asserting that probability measures would be taken into consideration with the rest of the available evidence in assessing the strength of a case (*Palmer v. Shultz,* 1987; *Vuyanich v. Republic National Bank of Dallas,* 1981).

Quite clearly, because the employment practices being examined in disparate-impact cases are "fair in form" (*Griggs v. Duke Power Co.,* 1971, p. 431) and the only question is the outcome of their application, statistical evidence, albeit of an elementary nature, is at the heart of the case; no intent to discriminate need be shown. On the other hand, there has always been some controversy over whether statistical evidence can address the issue of disparate treatment, that is, not whether the existence of a standard was discriminatory but whether it was differentially applied. In disparate-treatment cases, it is necessary to show an intent to discriminate so that the question is whether such an intent can be inferred from statistical disparities. The Supreme Court has said that intent can be inferred (*Hazelwood,* 1977); but in fact, rarely has statistical evidence alone been found sufficient to prove discrimination. Even in *Griggs,* where it was not necessary to infer intent, there was an added factor of earlier intentional discrimination: prior to the passage of Title VII, African-Americans had been totally excluded from the lucrative line-repair jobs. Similarly, in a case where the imposition of the death penalty on an African-American defendant was being challenged, evidence was introduced to show that the "random" selection of jurors by drawing names from a fishbowl consistently resulted in all-white panels. Such a result was highly improbable, but the clinching factor was that the names of prospective white jurors were written on white slips of paper, whereas those of African-Americans were written on yellow slips (*Avery v. Georgia,* 1952; Gray, 1983).

As long as employment discrimination cases involved mostly blue-collar workers, courts were quite receptive to the statistical arguments presented by plaintiffs (see, for example, Finkelstein and Levin, 1990). Alleged discrimination in professional employment presents a more difficult situation (Bartholet, 1982). For example, courts have consistently accepted a Ph.D. requirement for college professors without applying the *Griggs* analysis. Qualifications for such jobs are frequently complex and, in particular, simple comparisons of mean salaries generally are no longer an appropriate measure of possible discrimination. Multiple regression modeling of salaries came into widespread use (Bergmann and Maxfield, 1975; Dailand et al., 1973; Finkelstein, 1980; Fisher, 1980; McCabe and Anderson, 1976; Scott, 1977); courts were already familiar with the methodology in other contexts such as economic analyses in antitrust. In *Bazemore v. Friday* (1986) the Supreme Court approved regression as a technique to identify salary discrimination.

THE STATISTICIAN AS ADVOCATE

Combining the professions of statistician and lawyer seems incongruous to many. The incompatibility they generally have in mind centers on their conception of statisticians as narrow, quantitatively oriented introverts and lawyers as extroverts who live in the world of words. As neither of these stereotypes is very accurate, it is not this dichotomy that is troubling. Rather it is that the role of a lawyer is as an advocate—within the confines of ethical conduct—to make the best possible case for her client, whereas the role of a statistician is not supposed to be. Even the best statisticians stumble; R. A. Fisher's advocacy of tobacco interests was an embarrassing culmination of a distinguished career.

Unlike lawyers, statisticians are not bound by an ethical code, although an ethics committee of the American Statistical Association has produced guidelines. The difficulties of enforcement even for those professions that have such codes suggest that formal rules of conduct for statisticians are not the solution to the problem of tailoring of statistics to conform to a point of view, whether that of the statistician or of her or his client (Fienberg, 1989; Fienberg, Krislov and Straf, 1988; Fisher, 1986; Gibbons, 1973; Meier, 1986). If one is convinced that judges or other policymakers for the most part only pay attention to the statistics that fit in with their own ideas, abuses by statisticians may not be troublesome. Nonetheless, the roles of statistician and advocate should be kept separate.

Statistical experts for opposing sides can present differing views without stepping across the bounds into advocacy or distorting the evidence; however, it

is not always easy to distinguish honest differences of opinion from inappropriate behavior. In a few cases courts have engaged statisticians to do independent analyses, but this practice, although often proposed, has not become widespread (Fienberg, 1989). Plaintiffs presented studies that used essentially the same regression models and obtained substantially similar results in *Melani v. Board of Higher Education of the City of New York* (1983) and in *Coser v. Moore* (1984). However, the judge in *Melani* found the results persuasive, whereas neither the trial judge nor the appellate panel in *Coser* was convinced that statistical evidence was adequate to prove the existence of discrimination.

FACULTY SALARIES

The study of faculty salaries for evidence of gender discrimination is an example of a problem with interesting aspects from both statistical and policy points of view. Over the years multiple regression analyses of faculty salaries have been done at hundreds of institutions, most not as a consequence of litigation. Having seen the high costs in time and money and the frequently unsatisfactory results (Gray, 1988; *Rajendar v. University of Minnesota,* 1984), many colleges and universities have chosen to attempt to resolve equity questions internally. Among the most complex of all employment contexts is that of college and university faculty. The reluctance of courts to find colleges and universities guilty of discrimination has been attributed to a failure to understand the complex peer-review structure of faculty employment or to misplaced deference to academic freedom (Bartholet, 1982; Gray, 1988). An alternative hypothesis is that courts understand all too well; they identify strongly with the decision makers in colleges and universities, for these are people much like themselves (Gray, 1988). As a result, judges, like many statisticians on the institutions' faculties, find it hard to believe the evidence of a pattern and practice of discrimination presented by statistics. There must, they believe, be some other explanation.

One court found no evidence of discrimination in a university's failure to tenure a woman it conceded was better qualified than her successful male colleague on the grounds that she had failed to obtain her colleagues' "esteem" (*Namenwirth v. Board of Regents of University of Wisconsin System,* 1985). Another court accepted the university's explanation that the candidate's field was not in the mainstream of her department's interest, even though she had been hired specifically because of her specialty (*Smith v. University of North Carolina,* 1980). Only three plaintiffs have ever succeeded in winning tenure in Title VII cases (*Ford v. Nicks,* 1989; *Kunda v. Muhlenberg College,* 1980; *Brown v. Trustees of Boston University,* 1990). On the other hand, the court in *EEOC v. McCarthy* (1985) found for the plaintiff, whose sole witness was a statistician expert. The plaintiff's statistics were quite striking, with probability values between 0.005 and 0.00001; on the other hand, the defendant used substantially the same model and found no statistically significant differences. The disparity was apparently due to the fact that in the defendant's model the values for prior experience were as determined by a college official rather than self-reported.

Discrimination in faculty salaries could be proved under either a disparate-impact or disparate-treatment analysis. However, courts have generally held that in order to apply a disparate-impact analysis, a particular employment practice that causes the impact must be identified (*Ward's Cove Packing Co. v. Atonio,* 1989), although the practice could be subjective (*Watson v. Fort Worth Bank and Trust,* 1988). Given the complexity of salary setting in most colleges and universities, doing so is generally difficult, although the appellate court in *Sobel v. Yeshiva University* (1988) recognized the disparate impact of a "guideline" process of determining raises.

Looking simply at mean salaries for faculty men and women does not tell us very much. It is generally agreed, for example, that education and experience can and should influence the salaries of faculty members, although exactly how these factors are to be measured or how they affect the determination of salaries may be the subject of considerable debate [*Denny v. Westfield State College,* 1987; *EEOC v. McCarthy,* 1985; *Penk v. Oregon State Board of Higher Education,* 1987; *Segar v. Smith,* 1984 (see *Segar v. Civiletti,* 1981); *Sobel v. Yeshiva University,* 1988; *Trout v. Lehman,* 1986 (see *Trout v. Hildago,* 1981).] Thus, if the faculty men are on the average more experienced and a larger proportion of them have terminal degrees, a difference in mean salaries may be a legitimate reflection of this difference in qualifications. Because we want to control for these and other variables, the obvious technique to use is regression, although agreement on its appropriateness is not unanimous (Barnett, 1982; Conway and Roberts, 1986; Dempster, 1988; Norris, 1986; Roberts, 1979; Scott, 1977; Stacy and Holland, 1984). Matching procedures (*Craik v. Minnesota State University Board,* 1984; *Mecklenburg v. Montana Board of Regents of Higher Education,* 1976), "reverse" regression (Ash, 1986; Conway and Roberts 1986), an "urn" technique (*Sobel v. Yeshiva University,* 1988) and a cohort analysis have also been proposed (*Segar v. Smith,* 1984; *Trout v. Hildago,* 1981). Gastwirth (1988) discusses some of the considerations surrounding the use of the urn and cohort methodologies. The goal is to produce a regression equation which reflects as accurately as possible the salary process without introducing into

the model any factors that are themselves the source of bias.

The Population

Statistical analysis can be useful both in individual cases and in class actions (where the claims of a "named plaintiff" are taken as typical of those of a class). There are certain procedural requirements to be met to certify a class: numerosity, typicality, commonality and adequacy of representation. Generally, a class of faculty women can be certified for claims of salary discrimination, but occasionally defendant institutions are able to convince the court that because personnel decisions are initiated at the departmental level the commonality and typicality requirements cannot be met. Frequently the class is limited to certain schools or split into tenured and untenured faculty. Statistical and anecdotal information about the general situation is almost always admissible in an individual claim, although relief is possible only for the plaintiff.

In a class action, a salary analysis needs to be done for the class members, free of possible bias from the inclusion of nonclass members. Even if the class includes the entire university, defendants frequently seek to analyze separately each rank, each school or college, or even each department. This has a sound basis in the fact that the more homogeneous the group, the more likely it is that the salary process can be modeled accurately. Because of the effect of sample size on statistical significance, it is advantageous to the plaintiff to keep the population to be analyzed as large as possible, consistent with claims for reasonable homogeneity of the group. At the same time, it is frequently in the best interest of the defendant to eliminate as many people as possible from the analysis. It often happens that those eliminated are high-paid men and low-paid women, thus assisting the defense, but in any event simply reducing the size of the group being studied may make a large salary difference become statistically insignificant. Even if the difference remains statistically significant, defendants, and even courts, frequently "double discount" small sample sizes by claiming that even a substantial, statistically significant difference is meaningless in so small a sample. On the other hand, in large groups, although small differences may be statistically significant, the court may decide that they can be discounted as having no "practical" significance. Some courts have been willing to accept the combining of results by means of various statistical techniques (*Eastland v. TVA,* 1983; *Hogan v. Pierce,* 1983), but others have not (*Penk,* 1987). In *Penk,* combining the results for eight institutions under the common control of the Oregon State Board of Higher Education produced results showing a consistent pattern of statistically significant discrimination, even though the results at each institution were not statistically significant (Gastwirth, 1988, vol. 2). The court in *Capacci v. Katz & Besthoff* (1983) criticized disaggregation of the population into separate groups as an obvious and unfair attempt to reduce the statistical significance of differences (Gastwirth, 1989).

The question of the inclusion or exclusion of certain groups of faculty frequently arises in class certification or salary studies: part-time faculty, nontenure-track faculty, nonteaching professionals (generally librarians, counselors, research faculty) and clinical faculty. Part-time faculty present the problem of how equitably to convert their salaries to full-time equivalents. Part-time and nontenure-track faculty often have jobs confined to teaching with no service or research requirements, and thus their work may be sufficiently different from that of regular full-time faculty that it is inappropriate to include them in the study. Research faculty also have different job responsibilities and are frequently on soft money. Clinical faculty are paid less well than other faculty in law schools and better than other faculty in medical schools; in medical schools, payments from private practice associated with the university hospital present another complication. Indicia of the homogeneity required to group people together include whether they have faculty rank, whether decisions as to their salary, promotion and tenure are made in the same way, and whether they are all in the same bargaining unit if the institution has collective bargaining.

There are other groups of faculty whom one side or the other may want to exclude; for instance, defendants may want to eliminate those with named chairs, Nobel prizes, distinguished awards of one sort or another (usually high-paid men) or those in traditionally female disciplines such as nursing or home economics (usually low-paid women). Usually everyone can agree to eliminate deans, vice presidents, and provosts, but department chairs, directors, and associate and assistant deans with faculty appointments are more problematical. If there are identifiable additional stipends for such faculty, the stipends can be subtracted from their salaries for analysis purposes, but frequently the base salaries of these administrators are still substantially higher on the average—not only in the years they serve as administrators but in subsequent years. Defendants frequently propose including a variable for "prior administrative experience;" if, as is frequently the case, nearly all former administrators are male, the question of equitable access to administrative positions becomes an issue. Crucial also is that faculty who may subsequently have left the institution not be removed from the studies covering earlier years; otherwise bias may be introduced if such faculty have left because of perceived unfair treatment. In any event, the legal requirement is for nondiscrimination with respect to the actual employees in the time frame under consideration, re-

gardless of what may have happened to them subsequently.

Another procedure used to eliminate certain faculty from the group to be studied is to predict salaries by using a regression model and then to single out for elimination those faculty whose actual salaries fall substantially above or below their predicted salaries. The argument is that these persons are unique in some way – superstars or nonproducers – and they are distorting the model. Clearly the removal of outliers will improve the predictive power of the regression model, but with the loss of inclusiveness. If the outliers are thought to be the result of misrecording data, then if the data cannot be corrected, eliminating the outliers makes sense. However, if the data are accurate, removing even those outliers that have a substantial effect on the regression model may mask illegal discrimination. For example, if one man is paid much more than a similarly situated woman, the fact that other men are paid the same as she is does not immunize the employer from liability. An attempt by the defense to remove outliers who were men with prior administrative experience was discredited (*Denny,* 1987), but the defense in *Penk* (1987) was more successful.

Before the *Bazemore* decision in 1986, faculty hired prior to the liability period in a particular case or prior to 1972, the effective date of Title VII, were often removed from salary studies. The theory was that, whereas the initial salaries may have been discriminatory, subsequently raises were determined in a nondiscriminatory fashion. The only discrimination, it followed, had occurred at a time such that the employers were no longer liable or even at a time when the discrimination was not illegal. An alternative to a study of the salaries of only those faculty hired during the liability period is a study of the changes from year to year in the salaries of all faculty combined with a study of the initial salaries of faculty hired during the liability period. The reasoning is that the only salary decisions actually made during the liability period were these. In a sense this is equivalent to studying hiring ("flow") rather than the overall composition of the work force ("stock").

This theory is based on *United Airlines, Inc., v. Evans* (1977). In *Evans* a stewardess was fired for violating a rule against marrying, a rule that applied only to women; she did not file a complaint at that time. Later she went back to work but lost the seniority she would have had but for the earlier discrimination. The Court said that her present disadvantaged situation was merely the continuing effect of time-barred discrimination, not presently actionable.

After *Evans* defendants and some courts were interpreting pay discrepancies that were based on preliability actions to be similarly protected. However, in *Bazemore* the Court said that each paycheck issued at a discriminatory salary rate constitutes a discriminatory act and that if salaries had in the past been discriminatory they must be adjusted to be equitable. It is possible to interpret *Bazemore* to necessitate a study of whether the faculty salary structure was discriminatory before the start of the liability period or to require that salaries of faculty hired in the preliability period be studied separately from the salaries of those hired later. Treating pre- and post-1972 hires separately may be helpful if the discrimination affects only one group, although if a separate study is not done for each group, the salary differentials may be obscured by more favorable treatment of the rest of the population (*Sobel v. Yeshiva University,* 1988). That the bottom line shows no overall discrimination is not a defense against discrimination against one group (*Connecticut v. Teal,* 1982).

Men Only versus Total Population

The underlying concept in the prohibition of discrimination is that no one should be treated differently because of gender, race, religion or national origin. In modeling salaries by using multiple regression, the question is whether we should expect women to be paid the same as men who have comparable qualifications or whether we should expect women to be paid the same as the average of men and women. The problem arises because the regression lines for men and women may have not only different intercepts but also different slopes; for example, a Ph.D. may be worth an additional $1,000 on the average for men but only $500 for women. Comparing a males-only model of salaries with a females-only model shows the different values assigned to same characteristics and can more clearly identify the sources of discrimination. Although it would seem clear that using a men-only model to predict the salaries of women makes more sense statistically and legally, in at least one case the court asserted that using a men-only model overstates the gender disparity (*Ottaviani v. State University of New York at New Paltz,* 1989; see also Gastwirth, 1989).

There are some problems with the men-only model. For example, certain categories may have few if any men – schools of nursing, the instructors' ranks – thus making it difficult to get a meaningful prediction for these subgroups. It could also be that just as women would be paid more if they were paid as males are, men would get more if they were paid as women are. Thus as a check it is necessary to construct a women-only model and to predict the salaries of the males, which may be difficult if there are few women faculty. If there actually is discrimination against women, one would expect a negative average residual in the case of predicting salaries of women from those of men and a positive average residual in the case of predicting salaries of men from those of women.

It can be that negative residuals result from both models, in which case there may be no gender-based inequities. For example, if nearly all the women faculty are relatively inexperienced, there may be a quite large experience coefficient in a women-only model. If the men are mainly more senior, very likely the experience coefficient for the men-only model will be smaller as salary growth generally slows down. If in addition the intercept is smaller for women, the salaries for the men will fall primarily below the regression line for the women-only model, and the salaries for women will fall generally below the regression line for the men-only model.

Another way to deal with the problem of different slopes as well as with other interactions among variables is to use interaction terms (Gastwirth, 1988). Thus, for example, one would include not only gender, education, experience and department variables but also variables for the interactions between gender and education, gender and experience, and gender and each department. If the population is relatively small, this proliferation of variables is less than desirable; that there should be five times as many observations as variables is an often cited "rule of thumb."

Measurement Errors

Productivity is what theoretically determines salary, but we have no way of measuring that directly. Thus we need to use proxies; measurement errors can result from both the omission of important proxies and the mismeasurement of those which are included. For example, experience may be "mismeasured" because we use a proxy for it (say age) or publications may be "mismeasured" because no consistent rule is followed as to what should qualify as a publication for the purpose of the study. In addition, the quality of the publications, the number and size of grants, and the effectiveness of teaching may be productivity variables that are not included at all. For the measured variables, say education, experience and number of publications, in many cases on the average, women have lower values than do men. We know that there may be variables that affect salary but are not included in the model; the validity of regression depends on the assumption that there is no gender-based bias in the distribution of the values of the omitted variables. If, however, men rank higher on the included variables, it is frequently asserted that it is then reasonable to suppose that on those variables which are excluded, say quality of publications, the values are not randomly distributed as the regression model assumes, but rather the women rank lower on these also. Thus there may be measurement errors in productivity. To the exent that these errors are biased in favor of women, they may cause women's salaries to be overpredicted. A proposed remedy is to use salary as the independent variable in a reverse regression as it can be measured without error. A serious problem is, of course, how to construct an appropriate single "productivity" dependent variable (Ash, 1986; Conway and Roberts, 1986; Roberts, 1979). Moreover, reverse regression may underestimate the discrimination against women (Dempster, 1988).

Adjustments

Raises are frequntly given in terms of percentages, not in terms of dollar amounts. Therefore, a regression model using the logarithm of salary as the dependent variable may provide a better fit. On the other hand, as faculty acquire more experience the rate of growth of their salary tends to slow; sometimes adding the square of an experience variable to the model will give a better fit.

Variables

Central to the specification of the regression model is the question of which variables to include. The Equal Pay Act permits differences in pay based on seniority, quantity or quality of production, or a merit system, although it is not a sufficient defense merely to assert the existence of a merit system (*Marshall v. Georgia Southwestern College,* 1985); the courts have generally assumed that the Equal Pay Act defenses are incorporated into Title VII as well. There is general agreement that productivity and experience are, or at least should be, the legitimate determinants of salary. However, as productivity cannot be directly measured, the problem is what to use as proxies. Faculty productivity consists of scholarship, teaching, and service. We begin by looking at each of these in turn.

Scholarship

Information on publications or other indications of scholarship (or the equivalent in the case of artists and performers) is rarely easily available; generally the sources are curriculum vitae prepared by faculty and possibly citation indices or reviews. Even if the vitae exist, they are frequently not up-to-date, and the way in which the information is presented is not uniform. Obviously, institutions could improve this situation if they chose; however, resume writing is an art, and evaluating what is presented is difficult even for an experienced peruser of this art form. Although courts have not found it very satisfactory (*Gutzwiller v. Fenik,* 1988; *Merrill v. Southern Methodist University,* 1986), simply counting publications is usually all that can be done. Books and articles may be counted separately, but it should be taken into account that in most cases books are highly valued in the social sciences and the humanities but less so in science; distinguishing

between refereed and unrefereed publication (or juried and nonjuried shows in the case of artists) is not always easy; whether to give full credit to jointly authored works may also be an issue. However, if reasonably accurate information is available, it is probably worthwhile to include it simply to refute allegations that productivity is being ignored.

Counting the number of citations to a faculty member's work is some measure of its quality, but there may be a certain amount of bias in citation counts. Quantifying views expressed in reviews seems hopeless. Some procedures, such as including a dummy variable for the receipt of "prestigious" awards, are complicated by the difficulties in agreeing on what awards qualify; addition to the regression of a variable for receipt of an internal award for productivity had little effect in *Denny* (1987). Institutions could agree on basic guidelines to produce more useful vitae, and departments could specify the ranking of journals and awards, although subjective judgments are open to manipulation. The basic problem is that most institutions do not themselves have any detailed specifications for setting salaries.

Grant procurement is another indication of scholarship, as well as of the faculty member's value to the institution. Again, a simple count of grants or of their dollar value may be inadequate; aside from the differing opportunities for getting grants in various disciplines, there is a question of whether it is the total dollar value or the overhead received by the institution that should be considered. An analogous measure of productivity for clinicians is the amount of private practice revenue they generate as the university also receives a certain portion. Whether or not faculty have "graduate faculty" status or some similar designation is sometimes taken as an indicator of research productivity, but the process for acquiring such designations is frequently ill-defined and somewhat suspect; a logistic regression could examine whether there is evidence of bias in the process of awarding graduate faculty status.

More important than all of the difficulties in measurement is the fact that it is not clear that indicators of scholarship make any difference in salaries. At major research institutions, the filtering process is sufficiently thorough, particularly for tenured faculty, that the faculty may be so homogeneous that there are few significant differences in their publication records (Scott, 1977). At less research oriented institutions there may be so few publications on average that their effect cannot be detected, or it may be that the institution does not in fact reward publication in spite of its ostensible criteria for faculty productivity. For example, in *Ottaviani* (1989) the regression coefficients measuring the effect of publications in various years and various models ranged from a marginally significant −$27 to a lesser insignificant positive amount.

Teaching

Most institutions simply have no usable data on teaching effectiveness. At best there may be student evaluations for some courses or records of awards for outstanding teaching. Little is known about the effect of quantity or quality of teaching on salaries. The assertion is often made that heavier teaching loads hamper the ability of women faculty to publish extensively. However, information on teaching loads is rarely sufficiently well documented to be included in a regression model. Moreover, whereas it is illegal to pay women and men differently for the same work, whether it is illegal to pay women and men the same for different work is not so clear (*Berry v. Board of Supervisors, Louisiana State University,* 1986; *County of Washington v. Gunther,* 1981; *Hein v. Oregon College of Education,* 1983). Thus there may be nothing illegal in assigning heavier loads to women faculty, although the morality certainly is questionable if they are at the same time held to the same research requirements as men.

Service

Vitae may shed some light on the service question, but little is likely to be illuminated. Attempts have been made to account for doctoral dissertation supervision and service on committees, but there is little indication that in fact these variables might influence salaries (Scott, 1977).

Rank

Given the demonstrated difficulties with measures of productivity, the easiest solution would appear to be to use rank as a proxy for productivity. In most institutions there is some sort of process that purports to examine productivity in deciding on initial rank and subsequent promotions. It is clear that rank is almost always a very good predictor of salary. It is also clear that at nearly all institutions the process for determining rank is much like the process for determining salary, so that if bias has infected one, it may well have infected the other. Combining this with the fact that at almost all institutions women are clustered at the lower ranks leads to rank being a "suspect" variable in a model designed to detect gender discrimination. The dilemma is that if rank is not used in a regression model of salaries, differences perceived to be based on gender may just be the result of differences in rank distribution between men and women; but on the other hand, the differences in rank distribution may be the result of discrimination. Thus discrimination in salary

could have two components—within-rank salary discrimination, and discrimination resulting from placement of equally qualified men and women in different ranks.

The court in *Chang v. University of Rhode Island* (1985) recognized this possibility and in fact described the defendant's study of salaries at hire as a "farrago structured so as to conceal possible discrimination." The defendant's expert used separate regressions for each rank; he also used only a proxy for experience and actual salaries, whereas the plaintiff's expert used actual experience and log salary, approaches preferred by the court. The differing results are shown in Table 1. The defendant's expert further disaggregated his study, presumably to show that some women were treated even more favorably than men, but in fact it may well have shown that the source of discrimination was that experienced females were being hired at too low a rank.

Courts have split on whether or not rank is an appropriate variable to include in a regression analysis of salaries (*Chang,* 1985; *Coser,* 1984; *Craik,* 1984; *Mecklenburg,* 1976; *Melani,* 1983; *Ottaviani,* 1989; *Penk,* 1987; *Presseisen v. Swarthmore College,* 1978; *Sobel,* 1988). More generally, the question is whether any variable which is under the control of the institution should legitimately be included in the regression model. Some would even argue that such measures as publications should also be excluded, reasoning that working conditions—teaching assignments, access to internal grants for research and travel—can be used to affect faculty members' research productivity.

TABLE 1
Chang v. University of Rhode Island salary at hire

Plaintiff's study
Variables
Independent: doctorate, years since degree, year of hire, departmental grouping, years and type of prior experience
Dependent: log of salary

Results

	Net effect of being female	Standard deviations	p (one-tail)
	−0.057	−2.268	0.012

Defendant's study
Variables
Independent: degree, years since degree, year of hire, departmental grouping
Dependent: salary

Results

	Net effect of being female	Standard deviations	p (one-tail)
Associate professor	−$135.90	−0.15	0.440
Assistant professor	−$412.20	−1.53	0.063
Instructor	+$ 30.20	+0.02	0.492

TABLE 2
Chang v. University of Rhode Island salary at hire

Defendant's study
Variables
Independent: doctorate, years since degree, year of hire, departmental grouping
Dependent: salary

Results

	Net effect of being female	Standard deviations	p (one-tail)
Inexperienced new hires			
Instructor	+$11.30	+0.02	0.492
Assistant professor	+$92.70	+0.28	0.390
Experienced new hires			
Instructor	+$1,037.30	+0.65	0.258
Assistant professor	+$ 454.50	+0.94	0.174

Note: It should be noted that for the regression for experienced instructors there were only 15 observations.

Unlike education, experience or field of expertise, rank is not itself a qualification; rather, it is, or should be, the result of qualifications. Thus one should be able to construct a model to judge whether rank, either initially or through promotions, has been assigned in a similar fashion to men and women of equal qualifications. The problem is that the same qualifications theoretically determine rank as determine salary so that if a model is to be constructed using rank instead of salary as the dependent (categorical) variable, the independent variables will usually be the same. If these variables really determine rank, then the inclusion of rank itself in the salary model is redundant; if they do not, the separate model with rank as the independent variable is not probative. What does result from the regression studies of rank is a decomposition of any gender differences into two parts: one represented by the disparity within ranks as demonstrated by a salary model, including rank as an independent variable, and the other represented by the gender difference resulting from differential assignment to ranks. The separate analyses of rank and salary are particularly problematical for the plaintiffs if the disparities are not statistically significant separately although the combined effect is.

Experience

How to measure experience is not the easy matter it may seem. Computerized records of defendant institutions normally show nothing more than date of hire at that institution. Because of the relative ease in acquiring this information and the fact that regression models generally show seniority to influence salaries, this factor is almost always included. If rank is being used, then seniority may be augmented or replaced by years in current rank or years in each rank; of course,

these variables suffer from the same possible taint of bias as does rank itself.

Accounting for prior experience is considerably more complicated. The fundamental problem is that information may not be available, or if it is, it may be of questionable reliability, coming either from individual faculty self-reporting on vitae or from information as to what the institution has credited as prior service in determining initial salary or rank (*EEOC v. McCarthy,* 1985). Faculty can be very creative in reporting their experience; at one institution a faculty member credited himself with two years of entrepreneurial experience for having had a paper-delivery route.

Whether to include all experience or only "relevant" experience, what experience is relevant, how to account for less-than-full-time experience, whether to use one variable for prior experience or to account for different kinds of experience (tenure-track college teaching, non-tenure-track college teaching, research, high school teaching, elementary teaching, business, military, etc.) are just some of the questions that arise. The court in *Chang* (1985) declared that prior experience was simply irrelevant.

Because of the inherent difficulties in measuring prior experience, frequently proxies are used. Available proxies are age, years since highest degree, years since other degrees, years between degrees or some combination of these. It may be the case that using age overstates women's experience and hence may attribute to gender salary differences that in fact are due to difference in experience (*Vuyanich,* 1981). However, in *Denny* (1987) and *Chang* (1985) it was found that replacing actual prior experience with age reduced rather than inflated the gender disparity.

Even if there is agreement as to how to account for the quantity of prior experience, questions remain as to how to evaluate the quality of this experience. Use of prior salaries as a measure of quality of experience in determining starting salaries has been condoned (*Kouba v. Allstate Insurance Co.,* 1982), but such information is rarely widely available. Use of prior rank or prior tenure status are also favored as indicators of quality of experience (*Coser,* 1984; *Merrill,* 1986; *Ottaviani,* 1989).

Education

Variables could include possession of a doctorate (with possible differentiation among Ph.D., D.B.A., Ed.D. and J.D.), possession of a terminal degree (M.F.A. in arts, for example), possession of a second doctorate, second master's or professional degree, A.B.D. status (all but dissertation), master's plus a certain number of hours. Occasionally a variable for quality of degree, based on the rankings of the graduate programs of the institutions from which the faculty members received their degrees, has been added (*Coser,* 1984). Generally, possession of higher degrees is correlated with higher salaries; however, if educational attainments are highly correlated with other variables in the model, particularly rank, this may not be the case (*Denny,* 1987). Qualifications such as being a C.P.A. or other certifications are also sometimes used as variables (*Sobel,* 1988).

Market factors

That women tend to be clustered in generally low-paying disciplines may explain some of the disparities between the salaries of men and women. There are various ways to account for these disciplinary variations, or "market factors," as they are usually described, although "market factors" sometimes refers to the bargaining power of an individual faculty member. In addition to basic salary differences, other qualifications may be differently valued in various disciplines; for example, a Ph.D. may be of little value in the performing and visual arts. If separate regression models for each department are used, then interactions would not matter. However, most departments are far too small for this to be a meaningful option. A modification would be to have separate models for each school or college, but there may still be too few people or conversely there may be insufficient homogeneity. Such disaggregation may be a necessity, however, if the character of a school is very different as in the case, for example, of medical schools. Separate models could be used even for relatively small units and the results aggregated, but courts may resist aggregation (*Penk,* 1987). Using dummy variables for departments or groups of departments is a generally accepted technique. However, "market factors" may be different even within a discipline; for example, applied mathematicians may be paid more than theoreticians.

"Market factors" do not necessarily behave as one might expect. For example, in *Penk* (1987) the women in computer science were paid less than the men in home economics. Moreover, there is some evidence that salary differences between men and women are greatest in those disciplines that are predominantly male (*Chang,* 1985; National Research Council, 1979). Another way to measure "market factors" is to use external data collected by the College and University Personnel Association (CUPA) or a group of peer institutions relating overall salaries to salaries in a particular discipline. For example, if faculty in economics are paid on average $60,000 and faculty overall $40,000 on average, 1.5 (or its natural logarithm) could be used as the value of a market factor variable in a regression model.

Other variables

Administrative experience of one sort or another is frequently proposed as a factor explaining salary

disparities. Inclusion of a dummy variable for prior administrative experience was approved in *Ottaviani* (1989), even though no woman faculty member had had such experience. The addition of a dummy variable for whether appointments are for 10 months or 12 months has been suggested because of the uncertainties about how equivalency is to be established (*Penk*, 1987). Tenure status is also sometimes used as a variable; inherent in its use are the same problems as are encountered in using rank. In *Ottaviani* (1989) the defendant's expert used a categorical variable for year of hire, purportedly to reflect budgetary difficulties, as well as to substitute for years of experience at the institution. In fact, the coefficients for year of hire were very volatile; the years characterized as bad budget years developed positive coefficients after only a couple of years.

Another variable that has been used (McCabe and Anderson, 1976) is a subjective evaluation by a department chair or rank and tenure committee. Clearly the use of such a measurement suffers from the same infirmities as does rank. A general consideration is that factors under the control of the employer may be subject to the same bias as the salary itself and thus may be inappropriate for inclusion in the regression model. Before the *Bazemore* (1986) decision, it was common to use initial salary as an explanatory factor for current salary; although the predictive value might be excellent, its inclusion means that all that can be measured by the model is the possible discrimination in raises, not in salary.

No matter how many factors are in the model, it is inevitable that some additional attribute will be alleged to account for apparently gender-based disparities. For example, in *Penk* (1987) even the defendant's expert testified that there was no reason to believe that the women faculty were less well-qualified than the men faculty, but the judge still found that some factor not included in the regression models might have explained the statistically significant differences between men's and women's salaries. As discussed above, there could be an attenuation effect, but in general, the courts' refusal to accept well-constructed statistical models as evidence of discrimination is not based on statistics but on an unwillingness to accept that statistics can prove what the decision maker does not want to believe. The dissent in the circuit court decision in *Watson v. Fort Worth Bank and Trust* (1988) pointed out quite clearly the failure of the majority properly to evaluate the statistical evidence (Gastwirth, 1989).

TABLE 3
Ottaviani v. State University of New York at New Paltz: Values of coefficient for year-of-hire variable in defendant's salary study

Year of study	Year hired					
	73-75	76	77	78	79	80 or later
1973	$1,258					
1974	441					
1975	728					
1976	628	65				
1977	486	−54	396			
1978	226	397	−83	1,462		
1979	−529	566	−862	783	68	
1980	586	2,009	613	416	1,379	4,450
1981	541	2,739	485	146	1,282	2,278
1982	1,488	2,362	1,467	1,359	1,294	4,390
1983	2,146	3,585	1,994	2,741	2,163	4,643
1984	2,522	4,355	1,831	3,512	467	6,412

TABLE 4
Ottaviani v. State University of New York at New Paltz plaintiff's regression

Variables: Doctorate, years since doctorate, current rank, years in rank, rank prior to being hired at SUNY New Paltz, department group, prior experience (college teaching, high school teaching, other), longevity at SUNY New Paltz

Academic year	Net effect of being female	Number of standard deviations	Probability (two-tail)
1973/74	−$462	−2.19	0.0286
1974/75	−772	−2.92	0.0034
1975/76	−694	−3.10	0.0020
1976/77	−862	−4.07	0.0001
1977/78	−867	−4.38	0.0001
1978/79	−391	−2.02	0.0434
1979/80	−509	−2.21	0.0270
1980/81	−727	−3.07	0.0022
1981/82	−1,195	−3.91	0.0001
1982/83	−1,284	−5.35	0.0001
1983/84	−1,225	−4.22	0.0001
1984/85	−645	−1.56	0.1188

Note: R^2 ranges from 0.75 to 0.91 (higher values in earlier years).

Table 4 shows the results of the plaintiff's basic regression model in *Ottaviani* (1989). A model excluding rank variables produced substantially larger net effects. Stipends for department chairs were subtracted from the salaries used in the study; however, the addition of a dummy variable for department chair reduced the disparity to statistical insignificance in 1978/1979 and 1979/1980. Publication variables were included in preliminary studies but were dropped when they were found to have negative insignificant effect. The results shown are for a men-only model with salary as the dependent variable for easy comparison; using log salary produced essentially unchanged results.

In Table 5 are displayed the results of the defendant's regression model in the same case. A total population model was used. The fact that the defendant's model did not account for experience consistently and added several arguably tainted variables did not keep the defendant from prevailing, possibly because of a failure of the plaintiff to make the statistical case

TABLE 5
Ottaviani v. State University of New York at New Paltz defendant's regression

Variables: Doctorate, years since doctorate, current rank, years in rank, rank prior to being hired at SUNY New Paltz, department group, prior experience (college teaching, other relevant), years at SUNY at New Paltz prior to 1973, year hired if hired after 1973, dummy variable for prior administrative experience at SUNY at New Paltz, dummy variable for department chair

Academic year	Net effect of being female	Number of standard deviations	Probability (two-tail)
1973/74	−$268	−1.06	0.2900
1974/75	−277	−1.07	0.2900
1975/76	−26	−0.11	0.9100
1976/77	−262	−1.09	0.2800
1977/78	−291	−1.13	0.2600
1978/79	−171	−0.62	0.5400
1979/80	−8	−0.02	0.9800
1980/81	−72	−0.21	0.8300
1981/82	−270	−0.58	0.5700
1982/83	−360	−0.63	0.5300
1983/84	+83	+0.14	0.8800

convincing or because of the weakness of the nonstatistical evidence or possibly because of the predisposition of the judge not to be convinced by statistics.

Other Statistical Considerations

It is possible to argue that a regression model is useful even with a low correlation coefficient; however, generally an adjusted R^2 of around 0.70 is necessary to convince the court to have some level of confidence in the model in the context of salary regressions (Fienberg, 1989; Finkelstein and Levin, 1990). Clearly, just looking at a summary measure like R^2 is not enough. Certainly the distribution of the residuals should be examined in order to check the appropriateness of the model.

It has also been suggested that changes in R^2 values due to the addition of a variable should be used as an indication of the effect of that variable on salaries, particularly in the case of the sex variable. However, the addition of a single variable is likely to have a small numerical effect in R^2, suggesting that the added variable is not important, no matter how large its actual effect on salaries might be (Finkelstein and Levin, 1990). Because of this tendency to minimize salary differences, the technique is more likely to be proposed by defendants than by plaintiffs.

In most regression models of faculty salaries there are many variables which are not independent of one another—for example, experience and rank or gender and education. Such collinearity can lead to instability in the values of the individual coefficients; removal or addition of a single data point to the model can cause substantial changes. If gender is not one of the variables in the model, this is not a significant problem because the total contribution from all of the collinear variables is relatively stable, and we are not concerned with the contribution from the individual variables. If gender is used as a variable, the existence of collinearity may produce such a large standard error in the gender coefficient that even large disparities will be statistically insignificant (*Denny,* 1987; Scott, 1977).

The general question of how to judge the statistical significance of disparities has been discussed above. An additional complication is the fact that regression models generally study salaries for a number of years. Clearly results from year to year are not independent of one another (*Sobel,* 1988); nonetheless, a consistent pattern would appear to bolster a charge that there is a pattern and practice of discrimination, not just isolated cases (*International Brotherhood of Teamsters v. United States,* 1977). Some courts have appeared to require that there be statistically significant differences each year in order to find any liability at all (*Ottaviani,* 1989), whereas others have not (*Craik,* 1984).

There remains the question of why we should concern ourselves at all about statistical significance, given that the whole population is being studied; either there is a disparity or there is not (Baldus and Cole, 1980). However, there are undoubtedly factors that influence salaries but are not included in the model that is used. The underlying assumption is that their distribution is random. Thus we can think of a particular set of faculty, their qualifications and their salaries as a sample of all possible arrangements. The probability thus measures the likelihood that the observed differential could be the result of chance. It is doubtful that courts actually subscribe to this analysis; in fact, they frequently assume that the distribution of the missing factors is *not* random while at the same time requiring high degrees of statistical significance (*Coser,* 1984; *Ottaviani,* 1989; *Penk,* 1987). From a complete disregard of the concept of statistical significance, many courts have come to an almost mystical reliance on *p*-values (Gray, 1983; *Palmer v. Shultz,* 1987).

Remedies

Once a model has been constructed, the question arises as to whether it should be used to determine if an individual faculty member is underpaid. It must be remembered that regression is a statistical technique rather than a prescription for setting salaries. There may very well be variables not included in the model but which nonetheless affect salaries; the model only presumes that these are not distributed disproportionately between males and females in such a way as to affect salaries on the average. Moreover, collinearity in the model may make individual coefficients unreliable as estimators. Thus application of the regression

model to determine individual salaries is problematical (Gray and Scott, 1980).

If the model indicates systematic discrimination, we would expect an effect on the salaries of women other than those whose actual salaries are below their predicted salaries. Moreover, if the regression model were to be used to adjust individual salaries, then it would seem only fair to reduce the salaries which exceed their predicted values, a most unlikely outcome. In addition, many men have salaries below those predicted by the model and might challenge the results of a process that remedies the salaries of their similarly situated female colleagues but leaves their own salaries untouched (*Board of Regents v. Dawes,* 1976; *Ende v. Board of Regents of Northern Illinois University,* 1983). Alternative remedies might provide for equal dollar amounts to each woman, for different dollar amounts depending upon rank or experience or some combination of these. Billard, Cooper and Kaluba (1991) call for rotating the regression line for women's salaries to coincide with that for men. Thus a woman's salary that was $500 below what was predicted by the regression equation for women's salaries would be adjusted to be $500 below her salary as predicted by the regression for men's salaries, and similarly for a woman who is "overpaid."

Inherent in any remedy will be a certain amount of what some will perceive as unfairness. However, if there are identifiable groups of women faculty (in a certain school, very recently hired, etc.), who have been demonstrably fairly treated, consideration must be given to excluding them from any systemic remedy. Nonetheless, there may be cases where as a result of the remedy women will end up being paid more than similarly qualified men and cases where the systemic remedy leaves women still being underpaid. Moreover, similarly qualified men will be paid varying amounts as will similarly qualified women. As President Carter said, "Life is unfair." There is no way to ensure absolutely equitable treatment of all faculty; all that the procedures described can guarantee is that gender is not the basis of the inequities.

CONCLUSION

The use of regression models to detect salary disparities is likely to continue. Because the issue is close to the personal interests of faculty and judges, the challenge to statisticians to make a convincing case for statistical evidence is great and in some instances may be insurmountable. In spite of the *Bazemore* admonition that the theoretical effects of unaccounted factors is insufficient to undercut evidence from a well-constructed model, there will always be a tendency to believe that some other factor must explain any disparity. That people much like the judges themselves are inclined to discriminate is not a welcome thought—especially if it suggests that decisions on professional advancement may not always be made based on merit. Even though statistics may not be able to tell decision makers what they do not want to hear, as always, statisticians are obliged to exercise their best skills and professional judgment. Careful attention must be given to refining underlying data, formulating the model, explaining to decision makers the role of statistical evidence and refraining from inappropriate advocacy. What is needed is to bring "the cold numbers convincingly to life" (*International Brotherhood of Teamsters v. United States,* 1977, p. 329). Attorneys do this by putting a "human face" on discrimination, getting a sympathetic plaintiff to relate her experiences. In the face of resistance, statisticians need to make the story told by the numbers equally understandable and compelling.

REFERENCES

Alabama v. United States (1962). 304 F.2d 583 (5th Cir.), *aff'd* 371 U.S. 37.

Allen v. Seidman (1989). 881 F.2d 375 (7th Cir. 1989).

Ash, A. S. (1986). The perverse logic of reverse regression. In *Statistical Methods in Discrimination Litigation* (D. H. Kaye and M. Aickin, eds.). Dekker, New York.

Avery v. Georgia (1952). 345 U.S. 559.

Baldus, D. C. (1990). *Equal Justice and the Death Penalty: A Legal and Empirical Analysis.* Northeastern Univ. Press, Boston.

Baldus, D. C. and Cole, J. W. L. (1980). *Statistical Proof of Discrimination.* Shepard's, Colorado Springs.

Barnett, D. (1982). An underestimated threat to multiple regression analyses used in job discrimination cases. *Industrial Relations Law Journal* 5 156-173.

Barringer, F. (1991). Commerce dept. declines to revise '90 census counts. *New York Times* July 16 A1, A16.

Bartholet, E. (1982). Application of Title VII to jobs in high places. *Harvard Law Review* 95 945-1027.

Bazemore v. Friday (1986). 478 U.S. 385.

Bergmann, B. R. and Maxfield, M. (1975). How to analyze the fairness of faculty women's salaries on your own campus. *AAUP Bulletin* 61 262-265.

Berry v. Board of Supervisors, Louisiana State University (1986). 715 F.2d 971 (5th Cir. 1983), 783 F.2d 1270 (5th Cir. 1986).

Billard, L., Cooper, T. R. and Kaluba, J. A. (1991). A statistical remedy to gender- and race-based salary inequities. Preprint, Univ. Georgia, Athens.

Blot, J. B., and Fraumeni, J. F. Jr. (1986). Passive smoking and lung cancer. *Journal of the National Cancer Institute* 77 997-1000.

Board of Regents v. Dawes (1976). 522 F.2d 380 (8th Cir. 1975), *cert. denied,* 424 U.S. 914.

Brown v. Trustees of Boston University (1990). 891 F.2d 337 (1st Cir. 1990).

Capacci v. Katz & Besthoff (1983). 525 F.Supp. 317 (E.D.La. 1981), *aff'd in part, rev'd in part,* 711 F.2d 647 (5th Cir. 1983).

Chakraborty, R. and Kidd, K. K. (1991). The utility of DNA typing in forensic work. *Science* 254 1735-1739.

Chang v. University of Rhode Island (1985). 606 F.Supp. 1161 (D.R.I. 1985).

Clark, M. J. and Grandy, J. (1984). *Sex Differences in the Academic Performance of Scholastic Aptitude Test Takers.*

College Board Report No. 84-8. College Board Publications, New York.

COHEN, J. E. (1990). DNA fingerprinting for forensic identification: Potential effects on data interpretation of subpopulation heterogeneity and band number variability. *American Journal of Human Genetics* **46** 358–368.

COLLEGE BOARD (1988). *National College-Bound Seniors: 1988 Profiles. Profiles of SAT and Achievement Test Takers. National Ethnic/Sex Profiles.* College Board Publications, New York.

COMMONWEALTH v. CURNIN (1991). 565 N.E.2d 440 (Sup.Jud.Ct. Mass. 1991).

CONNECTICUT v. TEAL (1982). 457 U.S. 440.

CONWAY, D. A. and ROBERTS, H. V. (1986). Regression analysis in employment discrimination cases. In *Statistics and the Law* (M. H. DeGroot, S. E. Fienberg and J. B. Kadane, eds.) 107–195. Wiley, New York.

COSER v. MOORE (1984). 687 F.Supp. 752, *aff'd,* 739 F.2d 746 (2d Cir. 1984).

COUNTY OF WASHINGTON v. GUNTHER (1981). 452 U.S. 161.

CRAIK v. MINNESOTA STATE UNIVERSITY BOARD (1984). 731 F.2d 465 (8th Cir. 1984).

DAILAND, M. G., DAWKINS, S. M., LOVASICH, J. L., SCOTT, E. L., SHERMAN, M. E. and WHIPPLE, J. L. (1973). Application of multivariate regression to studies of salary differences between men and women faculty. *Social Statistics Section Proceedings of the American Statistical Association* 120–130.

DEMPSTER, A. P. (1988). Employment discrimination and statistical science (with discussion). *Statist. Sci.* **3** 149–195.

DENNY v. WESTFIELD STATE COLLEGE (1987). 669 F.Supp. 1146 (D.Mass. 1987).

EASTLAND v. TVA (1983). 70 F.2d 63 (11th Cir. 1983).

EEOC v. MCCARTHY (1985). 578 F.Supp. 45 (D.Mass. 1983) *aff'd* 768 F.2d 1 (1985).

ENDE v. BOARD OF REGENTS OF NORTHERN ILLINOIS UNIVERSITY (1983). 563 F.Supp. 501 (N.D.Ill. 1983).

FIENBERG, S. E., ED. (1989). *The Evolving Role of Statistical Assessments as Evidence in Courts.* Springer, New York.

FIENBERG, S. E. (1990). An adjusted census in 1990? The judge rules and the PES begins. *Chance* **3** 33–36.

FIENBERG, S. E., KRISLOV, S. and STRAF, M. L. (1988). Statistics, expert witnesses, and the courts. *Chance* **1** 32–37.

FINKELSTEIN, M. O. (1980). The judicial reception of multiple regression studies in race and sex discrimination cases. *Columbia Law Review* **80** 737–781.

FINKELSTEIN, M. O. and LEVIN, B. A. (1990). *Statistics for Lawyers.* Springer, New York.

FISHER, F. M. (1980). Multiple regression in legal proceedings. *Columbia Law Review* **80** 702–736.

FISHER, F. M. (1986). Statisticians, econometricians, and adversary proceedings. *J. Amer. Statist. Assoc.* **81** 277–286.

FORD v. NICKS (1989). 866 F.2d 865 (6th Cir. 1989).

FREEDMAN, D. A. (1991). Adjusting the 1990 census. *Science* **252** 1233–1236.

GASTWIRTH, J. L. (1987). The statistical precision of medical screening procedures: Application to polygraph and AIDS antibodies test data (with discussion). *Statist. Sci.* **2** 213–238.

GASTWIRTH, J. L. (1988). *Statistical Reasoning in Law and Public Policy, Vols. I and II.* Academic, Boston.

GASTWIRTH, J. L. (1989). A clarification of some statistical issues in *Watson v. Fort Worth Bank and Trust. Jurimetrics Journal* **29** 267–285.

GIBBONS, J. D. (1973). A question of ethics. *Amer. Statist.* **27** 72–76.

GRAY, M. W. (1983). Statistics and the law. *Mathematics Magazine* **56** 67–81.

GRAY, M. W. (1988). Academic freedom and nondiscrimination: Enemies or allies? *Texas Law Review* **66** 1591–1615.

GRAY, M. W. and SCOTT, E. L. (1980). A statistical remedy for statistically identified discrimination. *Academe* **66** 174–181.

GREEN v. USX CORPORATION (1990). 896 F.2d 801. (3d Cir. 1990).

GRIGGS v. DUKE POWER Co. (1971). 401 U.S. 424.

GUTZWILLER v. FENIK (1988). 860 F.2d 1317 (6th Cir. 1988).

HAZELWOOD SCHOOL DISTRICT v. UNITED STATES (1977). 433 U.S. 299.

HEIN v. OREGON COLLEGE OF EDUCATION (1983). 718 F.2d 910 (9th Cir. 1983).

HOEFFEL, J. C. (1990). The dark side of DNA profiling: Unreliable scientific evidence meets the criminal defendant. *Stanford Law Review* **42** 465–492.

HOGAN v. PIERCE (1983). 31 Fed.Empl.Prac.Cases 115 (D.D.C. 1983).

INTERNATIONAL BROTHERHOOD OF TEAMSTERS v. UNITED STATES (1977). 431 U.S. 324.

KAYE, D. H. (1989). The probability of an ultimate issue: The strange cases of paternity testing. *Iowa Law Review* **75** 75.

KOUBA v. ALLSTATE INSURANCE Co. (1982). 691 F.2d 310 (9th Cir. 1982).

KUNDA v. MUHLENBERG COLLEGE (1980). 621 F.2d 532 (3d Cir. 1980).

LEWONTIN, R. C. and HARTL, D. L. (1991). Population genetics in forensic DNA typing. *Science* **254** 1745–1750.

MANTEL, N. (1990). What is the epidemiologic evidence for a passive smoking–lung cancer association? In *Proceedings of the International Conference on Indoor Air Quality, Tokyo, 1987* (H. Kasuga, ed.) 341–347. Springer, New York.

MARSHALL v. GEORGIA SOUTHWESTERN COLLEGE (1985). 489 F.Supp. 1322 (M.D. Ga. 1980), *aff'd sub nom. Brock v. Georgia Southwestern College,* 765 F.2d 1026 (11th Cir. 1985).

MCCABE, G. P., JR. and ANDERSON, V. L. (1976). Sex discrimination in faculty salaries: A method for detection and correction. In *Proceedings of the Social Statistics Section* 589–592. Amer. Statist. Assoc., Washington, D.C.

MCCLESKEY v. KEMP (1987). 481 U.S. 279.

MECKLENBURG v. MONTANA BOARD OF REGENTS OF HIGHER EDUCATION (1976). 13 Fair Employ. Prac. Cases (BNA) 462 (D.Mont. 1976).

MEIER, P. (1986). Damned liars and expert witnesses. *J. Amer. Statist. Assoc.* **81** 269–276.

MELANI v. BOARD OF HIGHER EDUCATION OF THE CITY OF NEW YORK (1983). 561 F.Supp. 769 (S.D.N.Y. 1983).

MERRILL v. SOUTHERN METHODIST UNIVERSITY (1986). 806 F.2d 600 (5th Cir. 1986).

NAMENWIRTH v. BOARD OF REGENTS OF UNIVERSITY OF WISCONSIN SYSTEM (1985). 769 F.2d 1235 (7th Cir. 1985).

NATIONAL COMMISSION ON TESTING AND PUBLIC POLICY. (1990). *From Gatekeeper to Gateway: Transforming Testing in America.* Boston College, Chestnut Hill, MA.

NATIONAL RESEARCH COUNCIL. (1979). *Climbing the Academic Ladder: Doctoral Women Scientists in Academe.* National Academy Press, Washington, D.C.

NEUFELD, P. J. and COLMAN, N. (1990). When science takes the witness stand. *Scientific American* **262**(5) 46–53.

NORRIS, B. A. (1986). A structural approach to evaluation of multiple regression analysis as used to prove employment discrimination: The plaintiff's answer to defense attacks of "missing factors" and "pre-act discrimination." *Law and Contemporary Problems* **49** 65.

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).

PALMER v. SCHULTZ (1987). 815 F.2d 84 (D.C.Cir. 1987).

PENK v. OREGON STATE BOARD OF HIGHER EDUCATION (1987). 826 F.2d 458 (9th Cir.), *cert. denied,* 108 S.Ct. 158.

PEOPLE v. CASTRO (1989). 545 N.Y.S.2d 985.
PRESSEISEN v. SWARTHMORE COLLEGE (1978). 442 F.Supp. 593 (E.D.Pa.), *aff'd without opinion,* 582 F.2d 1275 (3d Cir. 1978).
RAJENDAR v. UNIVERSITY OF MINNESOTA (1984). 730 F.2d 1110 (8th Cir. 1984).
ROBERTS, H. V. (1979). Harris Trust and Savings Bank: An analysis of employee compensation. Report 7946, Dept. Economics, Graduate School of Business, Univ. Chicago.
ROBERTS, L. (1991). Fight erupts over DNA fingerprinting. *Science* **254** 1721-1723.
SCOTT, E. L. (1977). *Higher Education Salary Evaluation Kit.* American Association of University Professors, Washington, D.C.
SEGAR v. CIVILETTI (1981). 508 F.Supp. 690 (D.D.C. 1981), *aff'd sub nom.* SEGAR v. SMITH, 738 F.2d 1249 (D.C.Cir. 1984), *cert. denied sub nom.* MEESE v. SEGAR, 471 U.S. 1115 (1985).
SHARIF v. NEW YORK STATE EDUCATION DEPARTMENT (1989). No. 88 Civ. 8435 (S.D.N.Y. 1989).
SMITH v. UNIVERSITY OF NORTH CAROLINA (1980). 632 F.2d 316 (4th Cir. 1980).
SOBEL v. YESHIVA UNIVERSITY (1988). 566 F.Supp. 1166 (S.D.N.Y. 1983), *rem.*, 707 F.2d 1478 (2d Cir. 1986), 656 F.Supp. 587 (S.D.N.Y. 1987), *rev'd and rem,* 839 F.2d 18 (1988).
STACY, D. R. and HOLLAND, C. L. JR. (1984). Legal and statistical problems in litigating sex-discrimination claims in higher education. *Journal of College and University Law* **11** 107-177.
TANDE, C. M. (1989). DNA typing: A new investigatory tool. *Duke Law Journal* 474-494.
TROUT v. HILDAGO (1981). 517 F.Supp. 873 (D.D.C. 1981), *aff'd sub nom.* TROUT v. LEHMAN, 702 F.2d 1094 (D.C.Cir. 1983), *vac. on other grds,* 460 U.S. 1056 (1983), 652 F.Supp. 144 (D.D.C. 1986).
UNITED AIRLINES, INC., v. EVANS (1977). 431 U.S. 553.
VUYANICH v. REPUBLIC NATIONAL BANK OF DALLAS (1981). 521 F.Supp. 656 (W.D.Tex. 1981), *vac. on other grds,* 723 F.2d 1195 (5th Cir.), *cert. denied,* 496 U.S. 1073 (1984).
WATSON v. FORT WORTH BANK AND TRUST (1988). 487 U.S. 977.
WARD'S COVE PACKING CO. v. ATONIO (1989). 57 U.S.L.W. 4583.
WOLTER, K. M. (1991). Accounting for America's uncounted and miscounted. *Science* **253** 12-15.
WOLTER, K. M. and CAUSEY, B.D. (1991). Evaluation of procedures for improving population estimates for small areas. *J. Amer. Statist. Assoc.* **86** 278-284.

Comment

Delores A. Conway

Gray's article addresses statistical problems and concerns prevalent in legal cases of employment discrimination. Although she focuses on universities and the academic environment, the statistical issues apply to more general employment settings. In particular, the treatment of outliers, omitted variables, measurement issues, selection of variables, delineation of the population and comparison across competing groups arise in most legal cases of employment discrimination (Finkelstein, 1980). These problems expose the heart of statistical evidence and determine its probative value in legal settings.

One of the strengths of the paper lies in the numerous citations to actual legal cases that illustrate the use of specific methods. Gray notes that similar statistical results may be probative in one case and completely dismissed in another. Tabulated results from two legal cases illustrate the interplay between the legal and statistical issues when assessing employment discrimination at universities.

I commend the author for a careful and comprehensive discussion of the legal and statistical issues. This paper should be especially useful to practitioners and provides a checklist of problems to be addressed in the development of statistical evidence. My comments attempt to clarify and extend the discussion, as well as provide an economic perspective. The multiple regression framework shows how the statistical issues are interrelated and how summaries change with different viewpoints of the data. Two examples from legal cases illustrate the statistical complexities in assessing discrimination across different job structures within an organization. We conclude with some additional comments on the role of statistical evidence in Title VII legal cases.

Delores A. Conway is Associate Professor of Statistics, School of Business Administration, University of Southern California, Los Angeles, California 90089-1421.

STATISTICAL MODELS AND JOB STRUCTURES

Gray lists many of the considerations in the use of statistical methods to measure discrimination. Although they are presented in a somewhat isolated fashion, many of the statistical difficulties are interrelated. Solutions to one set of problems often resolve or magnify others.

The development of appropriate statistical models for Title VII cases is not a simple matter, because of the lack of a clear, causal model of the employment process and of limitations from observational data.