Statistical Reasoning in the Legal Setting

Joseph L. Gastwirth

The American Statistician, Vol. 46, No. 1. (Feb., 1992), pp. 55-69.

Stable URL:
http://links.jstor.org/sici?sici=0003-1305%28199202%2946%3A1%3C55%3ASRITLS%3E2.0.CO%3B2-%23

The American Statistician is currently published by American Statistical Association.

Statistical Reasoning in the Legal Setting

JOSEPH L. GASTWIRTH*

Expert testimony, based on statistical data and inference, has become a routine part of legal proceedings. This article describes the role of statistics as evidence in a variety of cases and how the relevance of statistical exhibits relates to legal standards and burdens of proof.

KEY WORDS: Concepts of proof; Equal employment; Law; Statistical evidence; Statistical inference.

During the last 20 years statistical data have played an increasing evidentiary role in a wide variety of legal issues. Cases involving the fairness of tax assessments, the safety and efficacy of drugs and chemicals, charges of discrimination in employment and housing, the use of blood tests to determine paternity, the accuracy of forensic tests used to identify (or exclude) an accused person from blood, semen, or hair samples, trademark infringement, and compliance with the rules and regulations for government funding programs rely on statistical evidence. This article describes the role and use of statistical data and inference in legal proceedings.

Before examining data sets used in actual cases, it is important to discuss the nature of the judicial process and how it differs from scientific research. The next two sections are designed for this purpose. In Section 2 we introduce the concepts of proof and how the courts allocate the burden of proof between the parties. Because, with the recent *Ward's Cove Packing Company et al. v. Frank Antonio et al.* (1989) case, the Supreme Court recently clarified (the dissenting opinion was that it changed) the burden of proof in equal employment (Title VII) cases, we give more attention to these cases. In Section 3 we illustrate how statistical experts interact with lawyers to develop statistical exhibits that elucidate the main facts and issues of a case. We will reanalyze data from legal opinions and give examples of statistical presentations that were properly focused and some that were not. As our interest is in the role of statistical evidence rather than the development of new techniques, the data sets will be relatively simple and of general interest. Since no data set is perfect, in Section 5 we briefly discuss methods of assessing the seriousness of potential flaws. Of special interest is the assessment of the possible effect an omitted (perhaps even unknown) factor might have on the ultimate inference.

The use of statistics in legal cases has been the subject of a number of recent books and articles. Quite often we will refer to Baldus and Cole (1980), Gastwirth (1988), and the Committee of National Statistics (CONS) report (Fienberg 1989). Other important books are Barnes (1983), Finkelstein (1978), De Groot, Fienberg, and Kadane (1986), and Eggleston (1983).

1. THE GOALS OF SCIENTIFIC RESEARCH AND THE LEGAL PROCESS

The law recognizes the usefulness and relevance of scientifically accepted facts. For many years courts used the *Frye v. United States* (1923) criterion for admitting scientific evidence in criminal trials, that is, the testimony should be based on knowledge and/or techniques generally accepted in the particular field. However, the objectives of science and law are not identical. In the next section we will see that the law uses different standards of proof for different types of cases, while scientists in any discipline adhere to a single standard.

The purpose of science is not only to describe the world but also to understand the mechanism generating our observations. Thus, scientists develop general theories that explain why certain phenomena are observed and that enable us to *predict* further results. A theory is corroborated when predictions derived from it are borne out.

In contrast, the law is concerned with resolving a particular dispute. While it contemplates applying a general rule or approach to the case at hand, it is less concerned with developing a general method applicable to a family of similar cases than in deciding the case fairly. Justice Holmes said "It is the merit of the common law that it decides the case first and determines the principle afterwards." The notion of independent verification or reproducibility that is a basic tenet of scientific research is not directly applicable in the legal setting. A court must reach a decision within a reasonable time after the trial; it cannot wait for further evidence nor can it conduct further studies on its own. In contrast, science often progresses in stages. Small pilot studies relying on less rigorous methods are used to suggest potential hypotheses that are then subjected to a more careful study.

The need for courts to decide promptly is illustrated in cases involving a potential harm to the public, such as a defective or accident-prone product or exposure to a potentially toxic chemical or drug. It is not in the public interest to allow continued exposure to such risks while corroborating studies are being carried out if early studies indicate a *serious* risk. Indeed, to encourage manufacturers to increase the safety of their products, evidence of a postaccident modification is not usually admissible in product liability cases, which examine whether the product was manufactured and tested in accord with the appropriate standards at the time it was made, what the plaintiff knew at the time the product

*Joseph L. Gastwirth is Professor of Statistics and Economics, Department of Statistics, George Washington University, Washington, DC 20052. Joseph L. Gastwirth's research was supported in part by National Science Foundation Grants SES-8909866 and SES-9001458. It is a pleasure to thank William Schucany and the referees for many helpful comments and suggestions.

was used, and the adequacy of any warning label. Two important cases are *Jackson v. Firestone Tire and Rubber Company* (1986), which reviewed the relevant time frames in negligence and strict liability cases, and *Adams v. Fuqua Industries* (1987), where a new trial was ordered because the defendant was not allowed to submit evidence concerning the state of knowledge at the time the product was manufactured.

Scientific research places virtually no limitation on the nature of data nor on the way it is collected or analyzed, provided that the procedures are well documented and are based on a scientifically plausible approach. (In some settings there are ethical constraints on the gathering of data, e.g., informed consent in clinical trials.) Unlike the *Frye* doctrine, science encourages novel experiments and theories because unexpected findings will be reexamined by independent researchers. Courts are not only concerned with the relevance of evidence but also need to ascertain that it was properly obtained (the "exclusionary rule" prohibits government prosecutors from using evidence in criminal trials that was obtained in deliberate violation of the rights of the accused) and that it will be given its proper weight by a jury or judge. This last consideration has led to courts not admitting scientific testimony on the grounds that it will be given too much weight by a jury and thus may be prejudicial. Egesdal (1986) provided a nice survey of how courts examine scientific evidence and reviewed the psychological literature on how jurors react to both the message (actual words used by an expert) and the paramessage (the parts of the testimony contributing to an aura of scientific infallibility, such as the experience and prestige of the expert). He quoted several studies showing that jurors are not overly awed by scientific evidence, provided that they are able to understand it. However, the potentially prejudicial paramessage may be relied on by jurors when they do not. This means that statisticians need to prepare their courtroom presentations carefully using graphical and other more comprehensible techniques such as matching, in addition to fitting complex models to data; see, for example, Hoffman and Quade (1983).

Another aspect of the legal system that contrasts with science is the adversarial nature of legal proceedings. Each side is supposed to present its version of the facts from a view beneficial to it. Lawyers are not obligated to introduce evidence contradicting or questioning other evidence they submitted. It is the other party's duty to produce such countervailing evidence. Scientists, however, are expected to cite studies that disagree with or suggest limits to their findings. Meier (1986) provided a valuable discussion of the effect the adversarial process may have on an expert's objectivity, and the views of experts involved in the politically sensitive school desegregation cases are described in Chesler, Sanders and Kalmuss (1989). Another disquieting characteristic of a trial is that the lawyer need not provide an expert with all the available data. Thus, in a discrimination case, one side's expert may analyze payroll records while the other side's expert may work with information taken from application forms with the initial pay received. The main issue facing the judge may be which data base is more relevant and reliable, not which statistical analysis was better. Both experts may have carried out an appropriate analysis of the data they were provided.

Although there are differences in the purposes of science and law, as well as differences in their approaches to obtaining and evaluating evidence, there are also basic similarities. Only after there are a series of opinions on a common subject does the law need to reconcile the cases and deduce an underlying rule. Scientists may conduct a number of experiments in an area before a unifying theory emerges. More fundamentally, both fields have "ideals" they strive to meet. Science has the goal of understanding nature and develops theories that explain observed phenomena. The law is based on precepts such as fairness and equality and strives to create a society in which no one will gain by harming another and where citizens can rely on a consistent and predictable legal system. It needs to adapt these basic ideals to a constantly changing world and to balance individual rights with the needs of society as a whole (Cardoza 1924).

2. STANDARDS OF PROOF USED IN THE LEGAL PROCESS

Although standards of proof, as well as the basic paradigm within which scientists operate, are subject to significant upheavals (Kuhn 1970), at any one time there usually is a single standard of proof that is accepted in any branch of science. The law, on the other hand, uses different standards of proof for different types of cases; those cases involving the most severe penalties tend to require a greater degree of proof. Although these legal standards are phrased in probabilistic sounding terminology, they are only roughly translatable into numerical quantities. In this section we review the concepts of proof used by courts and indicate how they interact with the rules for presentation of evidence in a trial.

For a person to be convicted of a crime, the prosecutor has to convince a jury that the accused is guilty "beyond a reasonable doubt." In most civil cases, which only involve monetary claims, the judge or jury base their decision on the "preponderance" of the evidence. I next list four definitions of evidentiary standards used by courts and ask the reader to assign probabilities to them. Note that these probabilities are conditional, as they refer to the fact finder's assessment of the evidence at the conclusion of the trial.

The legal standards are:

1. preponderance of the evidence or "more likely than not"
2. clear and convincing evidence
3. clear, unequivocal, and convincing
4. beyond a reasonable doubt

In deciding *United States v. Fatico* (1978), Judge Weinstein asked his colleagues to give numerical values to each standard. We reproduce the results in Table 1. How do your probabilities compare with those given by

Table 1. Probabilities Associated With the Various Standards of Proof by the Judges in the Eastern District of New York

Judge	Preponderance (%)	Clear and convincing (%)	Clear, unequivocal, and convincing (%)	Beyond a reasonable doubt (%)
1	50+	60–70	65–75	80
2	50+	67	70	76
3	50+	60	70	85
4	51	65	67	90
5	50+	Standard is elusive and unhelpful		90
6	50+	70+	70+	85
7	50+	70+	80+	95
8	50.1	75	75	85
9	50+	60	90	85
10	51	Cannot estimate numerically		

Source: *United States v. Fatico* 458 F. Supp. 388 (1978) at 410.

the judges? While there is essential agreement on the meaning of "preponderance" of the evidence, the range of the probabilities attached to "beyond a reasonable doubt" is somewhat disturbing. When I ask statisticians about this criterion, almost all responses are .95 or greater and very few are less than .9. Table 1 has been discussed by Solomon (1982), Gastwirth (1988, sec. 11.4) and Fienberg (1989), so we simply note that, while the judges agreed on the order of the strength or degree of proof required by the various standards, only the preponderance criterion was consistently translated into a numerical probability. Each of the others had a range of at least .15, a meaningful difference where probabilities are concerned.

The preceding discussion should not be interpreted as a criticism of the legal standards, rather it illustrates the difficulty of translating legal criteria into a precise mathematical framework. These are so formidable that judges refrain from defining the standard of proof for juries [see Posner (1990, p. 212) for citations]. Furthermore, if one desires to study the legal decision making process in criminal trials, one needs to introduce a prior probability of guilt to calculate $P(G \mid E)$, where G denotes guilt and E the evidence. How does one translate the statement that "a person is innocent until proven guilty" into a precise prior probability? The role of Bayesian inference in judicial decision making has received substantial attention. We refer the interested reader to Kaye (1987a,b) and Tillers and Green (1988) for articles dealing with probabilistic evaluation of evidence and to Thompson and Schuman (1987) and Faigman and Baglioni (1988) for experiments demonstrating that people have difficulty in interpreting the Bayesian reasoning in criminal cases.

2.1 Tort Cases

When reading legal opinions that use statistical evidence, it is important to know both the type of case and the precise phase of the legal process the opinion concerns. For example, clinical trials and case control studies are used in tort cases that concern medical malpractice, harm done to an individual from exposure to a toxic agent, and failure to warn of a potential danger, as well as in the documentation of worker compensation claims. Because workers who accept benefits under special compensation laws often forego the right to sue for larger amounts, and as these laws are of a humanitarian nature [see *O'Keefe v. Smith Associates* (1965) at 362], the amount of evidence required to show "causation" is less than in a tort case. As several worker compensation cases are discussed elsewhere (Gastwirth 1988, sec. 14.2), we just note that epidemiologic data showing that a condition at work, usually exposure to a chemical, *increased* one's chances of contracting the disease is sufficient to establish a worker's eligibility for compensation. In tort cases, when one asserts that their illness (C) was due to exposure (E), courts often require data showing that exposure to the agent at least *doubles* the risk. The logic supporting this criterion is that, if the relative risk, $R = P(C \mid E)/P(C \mid E^C)$, of exposure exceeds 2.0, then the additional cases resulting from exposure form the fraction $(R - 1)R^{-1}$, of all cases occurring in the exposed group, which exceeds .5; thus the "preponderance" or "more likely than not" standard of proof is met. Of course, this is a simplistic translation of the evidentiary standard, and the specific facts of a case also enter into a judicial decision, When a judge or jury is convinced that the plaintiff was healthier than average prior to exposure, they may decide in his or her favor even if the relative risk is less than 2.0, as in *Sulesky v. United States* (1980). On the other hand, a woman who drank heavily during pregnancy and then sued the liquor company for compensation for her baby's birth defects lost a jury trial recently. The jury apparently felt that the woman should have known of fetal-alcohol syndrome and would not have changed her habit of heavy drinking even if a warning had been included on the label.

The conflict between scientific certainty and legal sufficiency arises often in toxic tort cases, especially those that concern a possible association between exposure and an illness that has not been studied extensively or where the measurement of exposure is subject to substantial error. Black (1988) discussed the *Frye* standard used to evaluate forensic evidence and praised tort cases that apply the same criteria in their evaluation of epidemiologic and medical evidence. Courts are, however, understandably reluctant to wait until a sufficiently larger number of deaths or cases of a serious disease occur so that a statistically significant relative risk can be established before they allow producers of the agent to be sued. The *Ferebee v. Chevron Chemical Company* (1984) opinion allowed a medical doctor to testify about causation even though no studies that focused on the particular illness had been carried out because the issue was on the frontier of science and the methodology used by the expert in forming his opinions was generally accepted. More courts are examining the reasoning and information underlying an expert's opinion and disregarding testimony that ignores contrary findings in published studies or relies on studies that did not yield statistically significant results. The recent opinion by Judge Garza in *Brock v. Merrill Dow* (1989) discussed the importance of the relative risk, R, and the relationship between statistical significance and the confidence interval (C.I.) for R; that is, if there is a true increased risk, the entire C.I. will exceed 1.0. The decision also stressed the preference of courts for experts who rely on peer reviewed studies and methodology.

2.2 Employment Discrimination Cases

The preponderance of the evidence standard of most civil suits also applies in employment discrimination (Title VII) cases. The Supreme Court has classified these cases into two types: disparate impact and disparate treatment. The first type concerns the effect of a specific employment practice such as administering a test or requiring a preset educational level. The second type is a general assertion of unfair treatment, which manifests itself in unequal pay or promotion rates for similarly qualified employees. Usually a greater degree of disparity is required in disparate treatment cases because the plaintiff needs to demonstrate that the employer intended to discriminate.

In Title VII litigation the plaintiff has the burden of ultimately persuading the court that he or she was discriminated against and has the initial burden of establishing a *prima facie* case. In disparate treatment cases this may be accomplished by showing that (a) the plaintiff belongs to a protected (minority) group; (b) the plaintiff was qualified for the job sought, and the job was available; (c) in spite of the plaintiff's qualifications, he or she was denied the job; and (d) the job remained open or was filled by a nonminority person with lesser qualifications.

Statistical data is most pertinent to the third aspect. Since it is difficult to evaluate an explanation of why a particular person was or was not hired, a statistically significant difference in the treatment of similarly qualified persons can be used by a plaintiff to strengthen his or her claim that he or she was denied the job for discriminatory reasons.

If the plaintiff establishes a *prima facie* case of disparate treatment, then the burden of production (of appropriate evidence) shifts to the employer, who needs to articulate a legitimate, nondiscriminatory reason for the complainant's rejection. When the employer presents such evidence, the plaintiff has the opportunity to show that the employer's explanation was a pretext. Of course, employers can use statistical evidence showing that the minority and majority group receive equal treatment (e.g., have the same hire rates) as part of their rebuttal of a *prima facie* case established on nonstatistical grounds. The *Texas Department of Community Affairs v. Burdine* (1981) opinion emphasized that the plaintiff's proof in the first and third phases must be by a preponderance of the evidence, while the defendant's burden of production only needs to raise a *genuine issue of fact* as to whether they discriminated.

The recent five-to-four decision in *Ward's Cove* (1989) makes the allocation of the burdens of proof and production in disparate impact cases more similar to disparate treatment cases than previously. The original disparate impact case, *Griggs v. Duke Power* (1971), concerned whether a requirement that applicants for blue-collar jobs have a high school diploma or pass a pen-and-pencil intelligence test had a legitimate business purpose or was used to discriminate. Before these new requirements were instituted, 10 years of school sufficed. The plaintiffs showed that, according to Census data, 34% of the White males in the state possessed a diploma, but only 12% of Black males did. This selection ratio of .353 clearly indicates that the requirement disproportionately excluded Blacks from the jobs. The Court ruled that such requirements have to be validated (i.e., shown to predict on-the-job success). The Court noted that the company had not submitted a study demonstrating a meaningful relationship of the requirement to on the job performance. Furthermore, during the first 18 months after the Civil Rights Act was in effect, White non-high school graduates had almost the same promotion rate as graduates, undercutting the defendant's claim that a high school diploma was job related. Requirements, such as the educational one in *Griggs*, having a selection ratio less than four-fifths are said to have a disparate impact on the minority group. Government guidelines require them to be validated.

In *Ward's Cove* (1989) the subjective employment criteria used by the Alaskan salmon canneries were alleged to have a disparate impact on minority employees. Unlike most disparate impact cases, which compared the "pass rates" of actual applicants or a proxy applicant pool derived from Census data, as in *Griggs*, plaintiffs compared the high percentage of non-Whites in cannery (unskilled) jobs to their low percentage in noncannery (skilled) jobs. Justice White's majority opinion deemed this comparison inappropriate. It stated that the proper basis for the initial stage of a disparate impact case is a comparison of the racial mix of employees (or hires) with their proportion of qualified persons in the labor market. The opinion reversed the Court of Appeals ruling that a *prima facie* case of disparate impact was established and remanded the case to the lower courts to determine whether such a *prima facie* case could be established on the basis of other evidence in the record.

Justice White's opinion continued, stating that,

> even if on remand respondents can show that nonwhites are underrepresented in the at-issue jobs in a manner that is acceptable under the standards set forth, this alone will

> *not* suffice to make out *prima facie* case of disparate impact. Respondents will also have to demonstrate that the disparity they complain of is the result of one or more of the employment practices that they are attacking here, specifically showing that each challenged practice has a significantly disparate impact on employment opportunities for whites and nonwhites.

If the plaintiffs meet this burden of proof, the employer will have the burden of producing evidence substantiating that the challenged practice significantly serves its legitimate goals. If the defendant produces evidence justifying the practice, the plaintiff may show that an alternative procedure for achieving the business purpose with less of a disparate impact on minorities exists, so the employer's justification is a pretext. Throughout the proceedings, however, the *burden of ultimate persuasion* belongs to the plaintiff.

The majority opinion was concerned about the fairness and cost of requiring employers to justify every statistical imbalance in its work force and felt that this would lead to the adoption of quotas. In contrast, Justice Stevens's dissenting opinion quotes Chief Justice Burger's opinion in *Griggs v. Duke Power* (1971):

> The objective of Congress in the enactment of Title VII is plain from the language of the statute. It was to achieve equality of employment opportunities and remove barriers that have operated in the past to favor an identifiable group of white employees over other employees. Under the Act, practices, procedures, or tests neutral on their face, and even neutral in terms of intent, cannot be maintained if they operate to 'freeze' the status quo of prior discriminatory employment practices.

Justice Stevens's opinion also notes that "The opinion in *Griggs* made it clear that a neutral practice that operates to exclude minorities is nevertheless lawful if it serves a valid business purpose."

The dissent emphasized the distinction between disparate treatment cases that concern whether an employer intentionally treated the employee unfairly because of minority status and disparate impact cases dealing with the effect of particular practices. It cites prior opinions stating that, once a disparate impact of an employment practice has been demonstrated, the employer shoulders the *burden of proving* the business necessity of the questioned practice rather than the lesser *burden of production*.

Comment. The opinion also raises a statistical issue, as each practice alleged to have a disparate impact will now be assessed *without regard* to the other practices. If one thinks of a three-stage hiring process in each stage of which the pass rate of the minority group is exactly four-fifths of the majority rate, then the total process has a selection ratio of $(.8)^3 = .512$; that is, an employer could use a set of practices leading to a minority pass rate just over one-half that of the majority, yet not one would need to be validated. This suggests that statistical methods should be developed for jointly testing the individual and total impact of various requirements.

Data from a variety of disparate impact cases is presented in Arvey (1979), Baldus and Cole (1987), and Gastwirth (1988). Although much of the discussion in the psychological and statistical literature on test-validation concerns the correlation between the job requirement and the main tasks inherent in the position, courts often use a decision-theoretic approach, accepting a lesser degree of correlation when the job involves risk to the public (bus drivers, police, etc.) than when more routine jobs are at issue.

In one of the first *post–Wards Cove* cases, *Allen and Battle v. Seidman* (1989), Judge Posner stated that in situations where there is a large disparity (39% vs. 84%) in pass rates and both groups are homogeneous with respect to relevant background factors, a simple comparison of the rates will suffice for plaintiffs to establish a *prima facie* case. The opinion noted that the defendant had the opportunity to rebut the plaintiff's evidence with a more sophisticated statistical analysis (e.g., logistic regression) incorporating other covariates. Furthermore, the judge observed that the reduced burden placed on a defendant who now needs to demonstrate a legitimate purpose, rather than business necessity, to rebut a plaintiff's *prima facie* case implicitly lessens the amount and strength of the evidence concerning the impact of a particular employment procedure that is needed to establish a *prima facie* case.

We end this subsection by illustrating how the standard of proof interacts with the shifting burdens of production and proof in an equal employment opportunity (EEO) case. In *Hopkins v. Price Waterhouse* (1989) the lower courts found that the plaintiff, who was denied a partnership in the firm, had demonstrated that her sex, along with other factors, played an important part in the employer's decision. The issue before the Supreme Court was the standard of evidence the employer would have to satisfy in demonstrating that the plaintiff would not have been made partner anyway due to other job-related factors. The Court of Appeals had said that the *clear and convincing* standard should apply in further proceedings. The Supreme Court reversed the appellate court's view, stating that the employer only needed to meet the *preponderance of the evidence* standard. While the plaintiff prevailed in placing the burden of proof, rather than production, on an employer, once she convinced the court that sex was a substantial factor the employer prevailed with respect to the "standard of proof." In *Hopkins v. Price Waterhouse* (1990) the district court reviewed the record under the preponderance standard and decided that Ms. Hopkins should be made a partner.

2.3 A Tax Assessment Case

We now describe a tax objection case, *Twin Manors Condominium Association v. Rosewell* (1988), in which *clear and convincing* evidence is required when a party asserts that they are bearing an unfair tax burden due to an improper assessment because it is presumed that the government's assessors perform their duty properly.

The condominium is located in Morton Grove, in Cook County, Illinois. The county is divided into 38 townships (Morton Grove is Maine township) that are grouped into four quadrants. Each year one of the quadrants is reassessed. Owner-occupied condominiums are treated as single-family residences, and real estate of the same class in any *county* should be assessed uniformly. Single-family homes are supposed to be assessed at 16% of fair market value. Twin Manors was assessed at 15.35% of its value; however, the Association submitted data showing that typical condominiums and single-family residences in Morton Grove and Maine township were assessed at 10.5% to 10.9% of their market value. On this basis, plaintiffs claimed that the assessments violated the statutory requirement of uniformity. We summarize the data submitted by the plaintiffs from sales records during the 1977–1981 period in Table 2. The fourth data set came from a state-sponsored study that reported assessment-to-sales (A/S) ratio data for each township.

The Illinois Court of Appeals upheld a lower court's finding that data showing the plaintiffs' property was assessed at a substantially higher rate than similar property in the township did not demonstrate disproportionate assessment under the *clear and convincing standard* because the law specifies that uniformity should be measured against *countywide* assessments.

The plaintiffs argued that Table 2 should be sufficient to rebut the presumption that the tax assessment was correct so that the tax collector should have the burden of showing that the assessments were fair. Moreover, the plaintiffs argued that the burden of proving discrimination on a countywide basis was too heavy, in part, because of the cost of a large-scale statistical analysis. The court noted that the law specifically refers to the county as the referent population and said that, if the plaintiff was unable to present a countywide study, at least it could have submitted ratios for some properties in other townships.

Comment. Although the Supreme Court will not review the case, the A/S ratio studies (Behrens 1977, Gastwirth 1982) conducted in the past by the Census Bureau would have clarified matters. The latest data (1982) for Cook County showed a median A/S ratio of 19.4%. Thus, the condominium does not seem to have a legitimate complaint if the county is the relevant population. On the other hand, with more than half the class 2 homes having an A/S ratio exceeding 19.4%, many properties in Cook County are being significantly overtaxed relative to the 16% standard. Unfortunately, the 1987 Census of Governments did not collect A/S ratio data, thereby making it more costly for citizens to obtain relief from unfair assessments.

3. THE USEFULNESS OF STATISTICAL EXPERTISE IN ASSISTING THE LEGAL PROCESS

3.1 Focusing on the Fundamental Issues of the Case

Courts need statistical expertise not just to calculate the result of a statistical procedure but to ensure that the methodology is appropriate for the data and that the analysis sheds light on a major legal issue. Sometimes simple statistical tables suffice to highlight the main point of a case. In other situations more complex analyses are needed. In almost all uses of data, the judicial process relies on expert testimony to assess the soundness of the data base and to properly interpret the results of a statistical analysis.

An interesting use of summary statistics occurred in a Canadian court. In late March 1982 law enforcement officers in Alberta received a report of moose poaching. The legal hunting season begins in late August and ends in late November. Some moose hair and bits of hide were found near some outbuildings in the area. The individual accused of poaching admitted he had killed the moose from which the hide had come but claimed he shot it during the regular season.

It is well known that winter ticks are likely to be found on deer and moose. They have a life cycle, starting out briefly (in September and October) as unengorged, (UL) and engorged (EL) larva, a longer period as unengorged (UN) and then engorged (EN) nymphs, and finally they become adults [classified as adult male (AM), unengorged adult female (UAF), and engorged adult female (EAF)] in the spring. Samuel (1988) described the life cycle of winter ticks and presented data on the percentage of ticks in each life stage by month, which we reproduce in Table 3. The data was based on 327,425 ticks taken from 66 moose during a seven-year period. To determine the time the moose was shot, all 875 ticks found on the piece of hide, which was frozen in the snow, were classified. Their corresponding percentages were: UL (0) EL (0), UN (26.2), EN (28.3, AM (31.2), UAF (14.1) and EAF (0). A parasitologist testified that this pattern indicated the moose was shot after February. Indeed, the actual percentages seem to fall between the March and April ones in Table 3 (except for the EAF category). This evidence played a major role in convicting the accused.

An inadequate statistical presentation may now be instructive. In *Cox v. Conrail* (1987) plaintiffs alleged that

Table 2. Assessment–Sales Data Introduced by the Condominium Association

Population	Median A/S ratio	Sample size	Number with A/S ratios of 16% or more
Morton Grove Condos	10.78		
Maine Township Class-2 (1981)	10.51		
Maine Township	10.88	459	9
Morton Grove Class-2	10.77	178	0
Twin Manors	15.35		

Source: The opinion, *Twin Manors Condominium Association v. Rosewell* 529 N.E. (1988) at 1106. Class-2 properties include condominium and single-family homes. A blank means the opinion did not report the data.

Table 3. Percent Total of the Three Life Stages of Winter Ticks on Moose From Alberta Between November and April, 1978–1986

Life stage*	November	December	January	February	March	April
UL	0	.34	0	.19	0	.01
EL	.03	2.93	0	0	0	.01
UN	98.39	95.9	85.55	65.32	35.44	12.07
EN	1.38	.35	10.19	16.07	23.06	10.68
AM	.14	.29	3.22	10.63	23.14	44.24
UAF	.05	.19	1.04	6.76	17.67	26.75
EAF	0	0	0	1.03	.69	6.24

*UL, unengorged larva; EL, engorged larva; UN, unengorged nymph; EN, engorged nymph; AM, Adult male; UAF, unengorged adult female; EAF, engorged adult female.

Conrail discriminated against Blacks and females in hiring and promotion. One aspect of the case concerned the fairness and job-relatedness of the test used to evaluate an engineer trainee's knowledge. To assess whether a test has a "disparate impact" on a protected group, courts typically rely on the standard test for a difference between proportions yielding a significant result at the .05 level and on whether the selection ratio meets the four-fifths rule.

The opinion notes that 533 Whites and 85 Blacks started in at least one of 46 training classes; 36 Whites and 15 Blacks left or were terminated for reasons other than their failure (twice) of the exam. Of those remaining, 87% of the Whites and 60% of the Blacks passed. The full 2 × 2 table is given in Table 4. Judge Flannery noted that, while the selection ratio (.69) was less than four-fifths, he would not accept legal counsels' calculation of the normal approximation, or Z statistic, which showed a statistically significant difference.

My calculation of the test statistic is

$$Z = \frac{P_1 - P_2}{\sqrt{P(1 - P)\,(1/n_1 + 1/n_2)}}$$

$$= \frac{.8692 - .600}{\sqrt{(.8325)(.1675)\,(1/497 + 1/70)}} = 5.65,$$

corresponding to a p value of less than .0001 so that, with a proper foundation and interpretation, the data and test should have been persuasive. The opinion points out that the court is not in a position to evaluate the appropriateness of the formula selected, how it should be applied to the data in the case, nor what the significance of the result is. It also questioned whether the sample size of 70 blacks was sufficient.

On appeal, plaintiffs argued that the trial judge's refusal to accept their Z statistic showing significance at the .05 level was reversible error. The appellate opinion, *Frazier v. Conrail* (1988), upheld the district court. It noted that plaintiffs decided not to introduce an expert to avoid an extensive exploration of statistical analysis as part of their trial strategy. The trial judge, therefore, could rely on his intuition in assessing the relevance and importance of the data. At first glance, the statistical reader may look at the data and calculated Z and wonder why courts desire expert guidance. We cannot expect courts to know that a sample of 70 is quite reasonable in these situations, that the normal approximation to the exact test is appropriate for the data in Table 4, and that a Z statistic of 5.6 standard deviations is highly significant (assuming that the groups being compared have similar qualifications). The fact that 60% of the Blacks passed the exam may have indicated to the judge that the test was not designed to exclude them. Finally, the data in Table 4 aggregated the results of 46 trainee classes. A statistician would examine the data to see whether there is evidence of a change in the pass rates over time or in the educational backgrounds of the trainees before calculating the statistic. Otherwise, a stratified analysis and/or logistic regression would be more appropriate. Had a greater proportion of the trainees dropped out of the program for "other reasons," the assumption of representativeness of the data in Table 4 might be questioned and the actual reasons for dropping out investigated.

Table 4. Pass–Fail Data, by Race, From Cox v. Conrail

Race	Pass	Fail	Total	Fraction
Black	42	28	70	.600
White	432	65	497	.869
Total	472	93	567	.833

Source: The opinion, 47 FEP Cases at 710.

This example shows that courts need statistical expertise to assure that the facts on which the legal process relies are valid. In turn, statisticians should be leery when asked by lawyers, often a few days before trial, to run a "routine" analysis for them on data they have compiled. Before testifying in court, we need to know what the data refers to, how it was collected and what fraction is missing or unusable in order to decide the appropriate procedure for analyzing the data.

Another plaintiff's exhibit from the *Conrail* case exemplifies how data can be so overinclusive that it fails to focus on the basic issues. In support of their hiring and promotion claims, plaintiffs introduced data for each year from 1977 to 1984 concerning the proportion of persons employed as engineers and firemen who were Black. While the proportion of Black engineers appears small, ranging from 1% in 1977 to 3% in 1984, *no* data giving the fraction of Black persons qualified for the job was submitted. Usually one considers the pool of applicants for the entry level job (fireman) and compares the hiring rates of the two groups. When applicant data is unavailable or unreliable (e.g., if applicants were obtained by word of mouth from a virtually all majority work force), a comparison group is constructed using Census data on persons residing in the labor market area who possess the skills required for the position [see Gastwirth (1981) and Baldus and Cole (1987) for further discussion]. Data on all employees includes those hired before the legally relevant time frame, fails to incorporate the fact that firemen must have at least a year's experience to be eligible for promotion, and confounds the hiring and promotion issues and ignores the voluntary quits. Because the data did not enable the court to compare the fraction of Black hires or promotions with applicants or a qualified labor pool, it was unpersuasive.

The *Conrail* case shows how the statistician can contribute to the process of translating the legal issue into a statistical framework. However, we need background information from the lawyer and other experts to create appropriate comparison groups, such as eligible employees. This process of translation involves determining the relevant population(s) to be studied, the parameters of interest and the statistical procedure to be used. Statisticians cannot determine what values of the parameter are legally meaningful, and sometimes the parameter of interest itself is legally determined. In disparate impact cases the government decided that the selection ratio p_1/p_2, not the odds ratio (Fleiss 1981), is the parameter of interest. Freedman (1985) discussed a tax assessment case concerning railroads in which the law specified that a weighted mean of assessment to market value ratios should be used although the median of these ratios is the commonly used measure.

3.2 How the Legal Context Affects the Meaning of a Parameter

To illustrate how the "legally meaningful values" of the same statistical parameter can vary widely with the legal application, we note that the fraction (π) of a population with a specified attribute occurs in several types of legal settings.

(a) The Food and Drug Administration (FDA) tests food and health products to determine whether the fraction of defective or harmful items is too large.

(b) Trademark infringement cases are concerned with the fraction of potential customers who are "confused" by a similar product, especially if its name or packaging resembles that of an established brand.

(c) In petitioning for a change of venue for a trial on grounds of prejudicial pretrial publicity, a survey of residents of the geographical area from which jurors will be selected can be used to estimate the fraction of persons who know about the crime or relevant events, as well as their views concerning the innocence or guilt of the accused. Of course, the actual venire from which jurors are chosen is subject to questioning about their knowledge of the case and any opinion they have formed about the case on *voir dire*.

Naturally, the critical value of π varies according to the type of case. In the classic *United States v. 43 1/2 Gross of Rubber Prophylactics* (1946) case, the FDA seized a shipment of condoms because 7.4% of a sample of them contained holes. The court noted that this percentage translated into about 1,650 defective items in the total shipment and would constitute a potential threat to public health in light of the purpose of the product (protection against syphilis). The court did not specify an allowable fraction of defective items; it accepted the FDA's determination that 7% constituted a health risk. Today, in light of AIDS, the FDA has established 4 in 1,000 as its allowable fraction of defective condoms in a shipment.

Courts have not established a minimum fraction of potential consumers who might be confused by a similar trademark. Elsewhere (Gastwirth 1988, chap. 9) citations to cases and the legal literature are given. Roughly speaking, courts consider that, if a substantial minority (25–30%) of potential customers might be misled, there is infringement. In *Brooks Shoe Company v. Suave Shoe Corporation* (1981) the defendant (Suave) rebutted a faulty survey by its own survey showing that only 2.7% of persons who purchased an athletic shoe in the previous year recognized the supposed unique mark of the plaintiff's brand. In determining whether infringement has occurred, courts consider other evidence, such as the price and frequency of purchase of the item, as well as evaluate the soundness of the survey.

Clearly, regular unfavorable news accounts and TV reports can create an atmosphere that makes it difficult for the accused to obtain a fair trial in the locale where the alleged crime occurred. In *Irwin v. Dowd* (1961) the Supreme Court reversed a verdict of guilt based on evidence that, of the *venire* of 430 persons from which the jury was chosen, 62% had a *prior* opinion of the defendant's guilt, while 90% had some leaning that way. Furthermore, two-thirds of the actual jurors who found Irwin guilty felt he was guilty *before* the trial. To avoid such situations courts try tot carefully question potential jurors about their exposure to stories about the crime and their feelings of guilt or innocence.

Surveys of the jury-eligible population are also used to demonstrate that the degree of pretrial publicity will make it difficult to find a fair jury. McConahey, Mullin, and Frederick (1977) described how such a survey was effective in helping convince the judge to change the venue of the Joan Little case. An attitudinal survey was made in Beaufort County, North Carolina, the scene of the murder of a jail attendant, a nearby county (Pitt), and a more urban county (Orange) farther from the scene.

In Table 5 we give a short summary of the results of the survey, in which we have simplified the questions for brevity. Although quite a high percentage of the jury-eligible population knew a lot about the case, the proportion who had strong prior feelings that Ms. Little was guilty was much smaller in Orange County. Moreover, the third question, designed to assess racist attitudes, indicated that the defendant might well face a prejudiced jury in the local area and a fairer jury could more easily be obtained in an urban county. The judge moved the trial to Wake County, which is demographically similar

Table 5. Percentage of Persons Responding Yes to Questions About Aspects of the Joan Little case

	County		
Question	Beaufort	Pitt	Orange
Have you heard "a lot" about the case?	76	76	78
Do you believe Ms. Little is guilty?	38	38	18
Do you believe Blacks are more violent than Whites?	63	64	35

Source: McConahay, Mullin, and Frederick (1977). About 150 persons in each county were surveyed.

to Orange County, and Ms. Little ultimately was acquitted. Notice that the survey used in this case not only showed a high degree of prior knowledge and belief of guilt in the local area but also showed that a fair jury could be found elsewhere in the state.

In *United States v. Haldeman* (1976) the "Watergate" defendants petitioned for mistrial, in part, because Judge Sirica did not grant a change of venue. The United States Court of Appeals said that the trial judge could place greater reliance on a comprehensive *voir dire* examination of potential jurors than on an *opinion poll* submitted by one side. The decision noted that the defendant asserted that 52% of the venire (presumably of several hundred persons) had some inclination toward guilt, while the government claimed that only about 36% had such a prior leaning. As only 14 (12 plus 2 alternates) impartial jurors need to be found, it was reasonable for the trial judge to use a careful *voir dire* procedure to select an impartial jury.

The defendants also introduced survey evidence reproduced in Table 6. Although the responses to question 3 concerning "feelings" do suggest that there was more prior feeling about the defendant's guilt in Washington, D.C., the responses to question 4, dealing with prior opinion, are more homogeneous. In his dissent, Judge Mackinnon focused on question 3, on which about 15% more jury-eligible residents of Washington, D.C., indicated the defendants were guilty, and quoted social science research showing that, after learning occurs, knowledge of details declines over time but the general ideas remain.

Comment. In its discussion of the case, the recent Committee on National Statistics (CONS) report (Fienberg 1989) noted that the presentation of the survey data

Table 6. Results of the Survey Concerning Potential Jurors' Knowledge of the "Watergate Affair"

Question	United States	District of Columbia	Delaware	Indianapolis Division Southern District of Indiana	Richmond Division Eastern District of Virginia
Have you ever read or heard anything about the fact that a number of President Nixon's former aides have been indicted for covering up the Watergate affair?					
Yes	91%	93%	97%	92%	88%
No	7%	6%	8%	6%	9%
Not Sure	1%	1%	—	1%	1%
No Opinion	1%	1%	—	8%	2%
Thinking of Nixon's former aides who are now under indictment—do you have an opinion on their guilt or innocence? [Response of those who knew of indictments.]					
Yes	68%	73%	64%	65%	60%
No	26%	15%	21%	33%	27%
Not Interested	1%	1%	1%	1%	1%
No Opinion/Don't Know	16%	11%	14%	12%	21%
How do you personally feel, do you feel they are guilty or innocent in the Watergate affair? [Response of those who knew of indictments.]					
Guilty	52%	67%	53%	52%	49%
Innocent	6%	2%	1%	2%	3%
Not Guilty Until Proven	24%	16%	21%	23%	19%
No Opinion/Don't Know	19%	14%	25%	23%	29%
Of those who knew about the indictment and indicated that they had an opinion, those opinions were as follows:					
Guilty	75%	84%	73%	79%	75%
Innocent	7%	2%	1%	8%	4%
Not Guilty Until Proven	15%	13%	20%	16%	18%
No Opinion/Don't Know	2%	1%	1%	2%	4%

Source: The opinion, *United States v. Haldeman* 559 F.2d (1976) at 178–179.

NOTE: Multiple answers were recorded in more than one category. As a consequence, the totals of all answers to these questions will exceed 100%.

was imprecise and not easily interpretable. The Committee also suggested that the survey may not have been given sufficient weight because the appellate court noted that opinion poll data "is open to a variety of errors." I believe that the court preferred to rely on information about the actual venire persons rather than on a survey of "registered voters," a population that does not exactly coincide with eligible jurors. Even if the voting rolls were the sole list from which jurors were chosen, at the time of the trial persons employed as teachers, law enforcement officials, doctors, and lawyers were typically excluded from jury duty. This may account for the lower percentage of potentially biased jurors on the venire than in the survey data. Surveys of the jury-eligible population are one step away from the actual venire, and courts find them less probative than surveys of the venire.

These change-of-venue cases illustrate why there is no set percentage of potentially biased jurors that will automatically trigger a change in venue. The statistics have to be interpreted in the context of the case, history of the region, and the size of the pool of potential jurors. Although a higher percentage (62.3%) of eligible jurors in Washington felt the accused were guilty than the percentage (38%) in the nearby counties who believed Ms. Little was guilty, the location of Ms. Little's trial was moved. Perhaps the history of discrimination in rural North Carolina plus the response to the third question in Table 5 indicated a deep seated hostility to Blacks. Unlike the questionnaire used in the *Little* case, the survey submitted by the Watergate defendants did not ascertain the strength of the feelings about the defendants' guilt or about the Nixon administration.

3.3 Methodological Issues

In this section we illustrate how courts have dealt with issues such as the significance and power of tests and combining results from stratified data. For expository purposes, the examples will lie closer to the extremes of judicial interpretation of statistics rather than the average. Therefore, it is important to remember that the statistical information given the judge is filtered through the lawyers. Almost every statistician I know who has testified in several trials can relate an incident in which he or she wanted the lawyer to ask further questions to clarify the meaning of the analysis or introduce more data, but the lawyer decided otherwise. While this section emphasizes statistical procedures, one needs to keep in mind the context of the case. In particular, the time sequence of events in an equal employment case or tort case involving a toxic agent is often more crucial than the method of analysis.

Although legal opinions sometimes confuse statistical significance or the p value with the probability of the null hypothesis being correct (Baldus and Cole 1980, sec. 9.4), courts have adapted reasonably well to the vague "two to three standard deviation" criterion for establishing a *prima facie* case of disparate treatment that the Court suggested in *Castenada v. Partida* (1977) and *Hazelwood School District v. United States* (1977). Indeed, Judge Higginbotham's discussion of the p value as a "sliding scale" indicating the strength of the evidence against the null hypothesis (*Vuyanich v. Republic National Bank* 1978) is well worth reading, as is Judge Posner's discussion of both statistical significance and the distinction between correlation and causation in *Tagatz v. Marquette Univ.* (1988). On the other hand, the power of a test to reject a meaningful alternative is rarely discussed (Fienberg and Straf 1982). In *Capaci v. Katz and Besthoff* (1981, 1983) the plaintiff attempted to prove discrimination in promotion by demonstrating that *two* female employees served longer than 24 males before being promoted using the Wilcoxon test. The defendant rebutted this significant result (p value = .02) by using the median test, which has *no* power when $m = 2$ and $n = 24$. The trial judge accepted the median test even though the defendant's expert admitted under cross-examination that the median test would not "find discrimination" if the difference in time to promotion were a million years. Because this data has been discussed by Hollander and Proschan (1979), Gastwirth and Wang (1987), and Finkelstein and Levin (1990), we will not reproduce it here. Another point, however, is worth noting. The original data referred to promotions from mid-1968 until January 1973, the date of the complaint. Three women were promoted later in 1973, which strongly suggests that qualified women were eligible for promotion *prior* to the charge. Neither the district nor the appellate courts focused on this apparent change in the defendant's practice.

Combination procedures, which are essential to extracting information from stratified data (Cochran 1954) and to interpreting the results of several similar studies, have sometimes been misused by experts, and courts may not realize that the analysis presented is incomplete. In its discussion of the use of statistics in environmental regulation the CONS Report (Fienberg 1989) presented an example of an expert misinterpreting the meaning of several epidemiologic studies, each of which found a nonsignificant risk of exposure to a chemical. The quoted statement uses "power" both in its technical and intuitive meanings. Moreover, due to the small sample sizes of these studies nonsignificant relative risks, exceeding 1.0, may well indicate a significant risk when a proper combination method is used. Some errors that occur when analyzing stratified data, such as employees in different locations, occupations, or seniority ranges, or survival rates from multisite clinical trials are:

1. overstratifying the data so that combination methods lose power
2. simply pooling several 2 × 2 tables into one large one (i.e., ignoring the strata)
3. using a procedure, such as Fisher's summary statistic ($-2 \Sigma \ln L_i$, where L_i is the p value of the test in the ith strata and which is designed for continuous data) on discrete or count data *without* making an appropriate correction
4. insisting on a statistically significant difference in each strata rather than using a proper combination technique

5. making tests in each strata and rejecting the null hypothesis in a stratum based solely on the p value in the stratum without considering the multiple comparison aspect

Courts often detect the first and second error but can hardly be expected to appreciate the technical nature of the third. The effect of discreteness on Fisher's procedure was investigated by Louv and Littell (1986), who combined binomial data sets. The data they used arose in *Cooper v. The University of Texas at Dallas* (1979) and concerned a change of sex discrimination in hiring faculty. In each of $k = 5$ divisions, the total number of hires (n_i) and the number (T_i) of females were reported. The T_i were compared to the number expected, $n_i P_{i0}$, derived from external data on the fraction (P_{i0}) of recent Ph.D.s in the disciplines who were female. Although the decision ultimately turned on the appropriateness of the availability fractions P_{0i} used, as hires at all levels were reported, we focus on the effect of discreteness. It is reasonable to test H_0: $P_i = P_{i0}$ for all i against H_0: $P_i < P_{i0}$ for at least one i.

The usual (uncorrected) Fisher test yielded a p value of .0773, while Lancaster's (1949) corrected method (Louv and Littell 1986) yielded a p value of .018. Considering that we are making a one-sided test and courts often require two-sided tests to yield significant results at about the .05 level, the difference in p values can be quite important. Incidentally, the MH procedure (Gastwirth 1984) based on $\Sigma\ (T_i - n_i P_{i0})$ yielded a one-sided p value of .026.

Although the fourth error occurs with some frequency in equal employment cases, statisticians know that insisting on significance at the .05 level in each of k strata implies that the overall level of significance is $(.05)^k$, rendering it virtually impossible to obtain a significant result. Rather than emphasizing mistakes, it is more illuminating to examine data from a recent case where Judge Gibbons intuitively combined the stratified data and formal methods confirmed her analysis.

In November 1983, a female employee of Shelby County Criminal Court filed a charge of discrimination in pay between similarly qualified male and female clerical workers. Subsequently, the equal employment opportunity commission (EEOC) took the case on behalf of 14 female clerks and filed a lawsuit in July 1985 after its efforts to conciliate the charge were unsuccessful. Although there were several different job titles, the judge first found that the basic tasks were the same. From longitudinal data on the salary of all 14 women and men hired at about the same time, plaintiffs grouped the data into four strata, defined by time of hire. In Table 7 we

Table 7. Pay Data for Male and Female Clerical Employees of Shelby County Criminal Court, Grouped by Time of Hire

			Salary in year				
Initials	*Sex*	*Hire date*	*1976 or initial*	*1978*	*1983*	*1985*	*1988*
Hired in 1973–1974							
F.R.	F	5/73	725	1,076	1,474	1,526	
J.P.V.	M	9/73	954	1,186	1,666	1,895	2,151
T.D.	F	1/74	844	1,024	1,403	1,527	1,702
C.H.	M	1/74	954	1,130	1,666	1,840	2,064
P.B.	F	5/74	700	975	1,403	1,532	1,767
L.A.	M	5/74	816	1,130	1,548	1,627	1,873
C.C.	M	5/74	816	1,130	1,548	1,697	
P.E.	F	8/74	600	975	1,403	1,537	1,780
G.V.	M	9/74	500	1,130	1,548	1,702	1,895
Hired in 1975–1976							
T.P.	F	5/75	525	844	1,112	*	1,543
G.L.	F	1/76	500	844	1,306	1,453	1,696
S.B.	F	2/76	550	952	1,336	1,601	1,827
D.V.	M	3/76	700	1,076	1,548	1,752	2,021
J.B.	F	9/76	576	952	1,336	1,468	1,686
Hired in 1978–1979							
B.W.	M	1/78	886		1,474	1,707	1,761
B.D.	F	9/78	458		1,000	1,095	1,231
B.P.	F	10/79	600		1,000	1,090	1,508
J.A.	M	10/79	600		1,157	1,298	1,600
Hired in 1982–1983							
F.D.	M	8/82	850		1,000	1,422	1,475
P.S.	F	9/82	800		929	1,116	1,372
M.D.	F	12/82	850		929	1,110	1,300
V.H.	F	1/83	850		929	1,110	1,283
S.C.	F	7/83	800		800	1,090	1,263

Source: Plaintiff's exhibit 3. The blank entries indicate that the employee left the office, and the asterisk denotes that Ms. P requested a reduced work day so her salary is not comparable to the others.

give an extract of the data that gives each employee's salary at the beginning of 1976 or their initial salary, if they were hired after 1976, and their final salary. As the author believes it is important to examine the situation at the time of the charge, we report the salary at the end of 1983, just after the charge was filed and at the end of 1985 after the lawsuit was filed. Salaries at the end of 1988 are also included because they reinforce the general pattern. The opinion noted that, in 1988, in every strata the highest-paid female clerk earned less than the lowest-paid male clerk. The decision then described the growth pattern of appropriate matched sets such as (*T.D. v. J.P.V. and G.V.*) and (*B.P. v. J.A.*). To assess the significance of the salary pattern observed by the judge, we can use the Van Elteren procedure for combining Wilcoxon rank tests in each group. Applying the normal form of the test (Gastwirth 1988, sec. 7.5) to the 1983 data yields a value of −3.52 standard deviations corresponding to a p value less than .001. Thus, the salary pattern is very unlikely to have occurred by chance, and the analysis supports the *prima facie* case of discrimination found by the judge.

In rebuttal, the defendant attempted to explain the salaries by questioning the seniority dates of a few clerks who started as transcribers and by questioning the performance of the clerks. The opinion noted that seniority could not explain the overall pattern because seniority was not strictly followed. Furthermore, while almost all employees performed well, the two weakest were males, who still earned more than comparable females.

To refute the relevance of the data in Table 8, the county attempted to minimize the role of seniority in determining pay by showing that there were many occasions when an employee received a raise that made his or her salary exceed that of an employee with greater seniority. The data is reported in Table 8.

Apparently, without the benefit of a statistical expert, Judge Gibbons realized that all the information is in the discordant pairs (M, F) and (F, M) and essentially used McNemar's sign test. We quote the opinion.

> Although defendants rely on evidence concerning occasions on which employees received raises over an employee with greater seniority, this evidence actually supports plaintiffs' position. On thirty-four occasions since 1976 a male employee has received an increase that resulted in his being paid more than a female employee with greater seniority. Yet on only eight occasions has a female employee received an increase that resulted in her being paid a salary higher than a male employee with greater seniority.

Calculation of the normal approximation to the sign test yields a Z score of −3.86, which is significant at the .01 level. Moreover, the ratio of (F, M) to (M, F) pairs in Table 8 estimates the ratio of the odds a female has of advancing over a more senior employee relative to those faced by a male. This odds ratio of 8/34 = .235 means that females had about *one-fourth* the odds males had of receiving a raise large enough to pass a more senior employee of the opposite sex. In other EEO cases, where the odds ratio was estimated from stratified data, Gastwirth (1984) found that courts tended to label odds ratios of about .70 as "statistically close." This is in rough agreement with the four-fifths rule as pass rates of .4 and .5 have a selection ratio of .8 and an odds ratio of .667. Thus, the odds ratio of .25 supports the judge's conclusion.

Table 8. Sex of Higher and Lower Salaried Employees When Salary Increases Resulted in Seniority Violations

Higher	Lower	
M	M	29
M	F	34
F	M	8
F	F	25

Source: The opinion, *EEOC v. Shelby County Government* 48 FEP Cases at 768.

Comments. The information obtained by examining salary raises that violate seniority is embodied in the salary data of Table 7. In every case where the pairs have initial or 1976 salaries that are "close" (e.g., P.E. and G.V. hired in 1974, B.P. and J.A. hired in 1979 or F.D. and P.S. hired in 1982), within a few years the male moved ahead of the female and the gap widens until the end of 1985. Between 1985 and 1988, the salary gap between the sexes remained about the same in the first two groups and narrowed in the last two. This apparent change in the growth of the gap makes mathematical modeling of the salary curves difficult. My experience with these cases suggests that the filing of the lawsuit by the EEOC in mid-1985 may underlie this observation. This is why I recommend (Gastwirth 1989) that data for a few years prior to the complaint be given more weight than data referring to the postcharge period in the initial determination of liability. Finally, the 1983 salary data has been reanalyzed by Bhattacharya (1989) using a nonparametric procedure for matched data. His results confirmed the finding of a significant "sex effect" and demonstrated that the salaries were not adequately modeled by a linear regression on seniority and sex.

Although most of the data sets we have discussed came from EEO cases, similar issues arise in cases involving medical data. After the FDA withdrew its approval of drugs that were supposedly effective for inflammations associated with dental and related medical procedures, the manufacturer sued. The firm asserted that the drug was effective because significance was found in six comparisons made on subgroups of patients. The appellate court opinion, *Warner-Lambert v. Hechler* (1986), did not accept this claim because 240 statistical tests were made. Assuming that the tests were carried out at the .05 level, even if some of the comparisons were dependent, one would expect 12 significant results due to chance. Thus, the data supports the FDA's assertion that the drug was ineffective. Again the main statistical issue is the proper combination of statistical procedures to account for the number of tests made.

4. ASSESSING THE SOUNDNESS OF AN INFERENCE

Almost all real-world studies of human subjects do not strictly satisfy the assumptions often made in the data analysis. Even in randomized clinical trials, the sample size may not be sufficiently large to ensure that the groups are balanced with respect to all relevant covariates (Altman 1985) and the standard methods of survival analysis (Cox and Oakes 1984; Breslow and Day 1988) assume dropouts leave at random. The validity of this assumption is difficult to check in studies of modest size. When one analyzes observational data, the available sample is self-selected, and information on potentially relevant covariates may be missing or of low quality. Without appropriate statistical adjustments these factors may have a substantial effect on one's inference. In this section we use Cornfield's approach to assess the potential effect of an omitted factor on the main conclusion of statistical evidence.

Cornfield's result (1959) concerns the strength of the association an omitted variable must have in order to explain a finding that exposure has a relative risk R. If a new agent (X) acts on the disease producing mechanism independently of exposure, then, to fully explain the observed risk R, two conditions must hold:

1. The relative risk, R_x of agent X, must be at least R.
2. The fraction of persons in the exposed group on whom the factor X operates or who possess it must be *at least* R times the corresponding fraction of the control group.

Cornfield et al. used this to conclude that it was highly unlikely that the smoking–lung cancer relationship ($R = 5$) could be explained by some other factor. Rosenbaum (1987) cited many recent references extending this idea to other situations and incorporating the sampling error inherent in the estimate of R or other measure of difference between the two groups. In using this methodology one needs to be careful. Just because it is possible for an observed association to be explained by a new factor does not imply that it is. Here statistics blends with knowledge of the subject matter; one needs to ascertain whether the major known covariates were controlled for in the analysis.

Sometimes a statistically significant disparity is "explained" by a new factor without the benefit of a stratified analysis demonstrating that the disparity is *fully* explained by the suggested factor. An example of this may have occurred in *Maloley v. Department of National Revenue, Canada,* described in Juriansz (1987). The Revenue Service administers an exam to persons applying for jobs as collection clerks. To obtain a "highly qualified" work force, the Service increased the passing score for applicants in 1984. Only 68 of 251 or 27.1% of the female applicants passed the test compared to 68 of 115 or 59.1% of the males. The standard test of equality of proportions yields a normalized difference of 5.8 standard deviations (p value $< .0001$), a highly significant result. When there is a significant disparity in pass rates, as in the United States, the defendant is required to produce evidence justifying the job-relevance of the test.

As part of its rebuttal the Revenue Service asserted that the two groups of applicants were not "equally qualified," as college graduates perform better on written tests and 52% of the men had college degrees while only 25% of the women did. The Appeals Board accepted this explanation and found that women were not discriminated against.

From a statistical viewpoint, the Revenue Service should have been asked to stratify the data by possession of a college degree to determine whether persons with a college degree really did have *twice* the pass rate of high school graduates. Cornfield's criteria state that, for a new factor (X) to explain a relative risk R (here $R = .271/.591 = 2.18$, which we round to 2), its relative risk, R_x, must be of a least R (2 here) and be R (two) times more prevalent in the more successful group. The Revenue Service's data showed that college education satisfied the second (prevalence) condition, but no data was submitted showing that college *doubled* one's chances of passing the test. Thus it is not clear that the Revenue Service completely rebutted the disparity in the success rates.

This type of analysis confirms the court's inference in *Allen and Battle v. Seidman,* discussed in Section 2.2. In that case, the defendant did not identify any factor (e.g., experience or special skill) that might double one's probability of doing well on the test and that was twice as prevalent among the White candidates.

In light of the U.S. Supreme Court's decision in *Wards Cove* emphasizing the importance of plaintiff's first making a more specific analysis to place a burden on the defendants of producing evidence *creating an issue of fact not of persuasion,* Cornfield et al.'s approach should aid courts in deciding whether an offered justification yields a full explanation of an observed disparity, based on a sound statistical analysis.

One way courts have assessed findings of statistical significance in EEO cases is to ask the experts to recalculate the test statistic assuming that one more minority person was hired or promoted. While this is a useful tool, the effect on the power of the test should also be considered.

There are many other important statistical problems such as missing data, misclassification and other measurement errors, and the suitability of proxy variables (e.g., using the location of a soldier's unit to indicate exposure to Agent Orange) that courts have faced. These topics could form an article by themselves, so we refer the reader to Dempster (1988) for their effect in EEO Cases, Gastwirth (1988, chaps. 13, 14) and Savitz (1988) for a discussion of statistical studies of health risks, and Weinstein (1988) for the views of the judge who handled the major Agent Orange case.

5. SUMMARY AND DISCUSSION

The types of data and evidentiary issues discussed in this article illustrate only a portion of the potential application of statistical and probabilistic reasoning in law. As in most uses of statistics, the most pertinent and effective analyses are intertwined with the subject matter. Thus epidemiologic studies need to incorporate the appropriate latency period to enable the issue of whether exposure to a particular agent causes or promotes a disease to be resolved. An excellent treatment of the scientific issues is Judge B. Jenkins's opinion in *Allen v. United States* (1984). Although *Allen* was overturned on technical legal grounds, at least two books have supported the moral correctness of the decision and Congress recently authorized compensation for persons affected by radiation from the A-bomb tests.

I believe statistical data will play an increasingly important role in clarifying the validity and relevance of the studies relied on by judges and policymakers when they decide product liability and toxic tort cases or place limits on exposure to chemicals. Bright, Kadane, and Nagin (1988), however, note that the cost-efficiency gains inherent in sampling have not been fully used in some tax cases, although samples and surveys are accepted in other types of cases.

The adoption of the "two to three standard deviations criteria" by the Supreme Court in *Castenada v. Partida* (1977) helped our profession gain importance in the public policy arena; however, hypothesis testing, especially when the type II error is not considered, is too limiting a view of inference. The complex problems of society, which reflect themselves in real-world cases, require a more decision-theoretic approach incorporating the costs to society, as well as to the litigants. This is especially important in deciding environmental, health, and safety issues, as well as the job relatedness of tests. Indeed, environmental decisions made now may have a greater effect on persons who are not alive today than on the generation making the decision—the group most affected may not even be a party to the litigation.

The use of statistical data in courts also brings new problems to the statistical community. It surely has contributed to the interest in measurement error and its effect on regression analyses [see Schafer (1987) and Dempster (1988) for references], increased the importance of carefully conducted retrospective studies (Lagakos, Wessen, and Zelen 1986), reminded us of the need for studies of power (Goldstein 1989), especially when the sample sizes are unequal (Gastwirth and Wang 1987), and of the need to carefully examine the assumptions on which our methods are based. The *Hopkins* case and toxic tort cases involving carcinogens highlight the need to estimate the relative importance of several factors that contributed to an outcome. The recent articles by Kruskal and Majors (1989) and Lagakos and Mosteller (1986) provide a good introduction to the statistical aspects of this issue. In sum, I hope this article demonstrates that statistical data and inference can contribute meaningfully to more reasoned legal and public policy decision making and that this area of application brings a wide variety of interesting data sets and new problems for us to analyze and solve.

ADDITION IN PROOFS

Since this article was accepted, the Civil Rights Act of 1991 was enacted. Section 105 reverses the part of the *Ward's Cove decision* that allowed defendants to justify a practice having a disparate impact by the "legitimate purpose" criteria. Such a practice now needs to be job related and consistent with "business necessity." Thus, the *Griggs* decision has essentially been restored. The first issue of the *Journal of the Royal Statistical Society*, Ser. A, for 1991 contained many interesting articles concerning applications of statistics in law and forensics.

[*Received October 1989. Revised October 1990.*]

REFERENCES

Altman, D. G. (1985), "Comparability of Randomized Groups," *The Statistician*, 34, 125–136.

Arvey, R. D. (1979), *Fairness in Selecting Employees*, Reading, MA: Addison-Wesley.

Baldus, D. C., and Cole, J. W. L. (1980), *Statistical Proof of Employment Discrimination*, Colorado Springs: Shepards/McGraw-Hill.

——— (1987), *1987 Cumulative Supplement to Proof of Discrimination*, Colorado Springs: Shepards/McGraw-Hill.

Barnes, D. W. (1983), *Statistics as Proof: Fundamentals of Quantitative Evidence*, Boston: Little, Brown.

Behrens, J. O. (1977), "Property Tax Administration: Use of Assessment Sales Ratios," *Journal of Educational Finance*, 3, 158–164.

Bhattacharya, P. K. (1989), "Estimation of Treatment Main Effect and Treatment-Covariate Interaction in Observational Studies Using Band Width Matching," Technical Report, University of California at Davis, Division of Statistics.

Black, B. (1988), "Evolving Legal Standards of Proof of the Admissibility of Scientific Standards," *Science*, 239, 1508–1512.

Breslow, N. E., and Day, N. E. (1988), *Statistical Methods in Cancer Research: Vol. II–The Design and Analysis of Cohort Studies*, Lyon: International Agency for Cancer Research.

Bright, J. C., Jr., Kadane, J. B., and Nagin, D. S. (1988), "Statistical Sampling in Tax Audits," *Law and Social Inquiry*, 13, 305–333.

Cardoza, B. (1924), *The Growth of the Law*, New Haven, CT: Yale University Press.

Chesler, M. A., Sanders, J., and Kalmus, P. S. (1989), *Social Science in Court: Mobilizing Experts in the School Desegregation Cases*, Madison: University of Wisconsin Press

Cochran, W. G. (1954), "Some Methods for Strengthening the Common χ^2 Tests," *Biometrics*, 10, 417–451.

Cornfield, J. C., Haenszel, W., Hammond, E. C. Lilienfeld, A. M., Shimkin, M. B., and Wynder, E. L. (1959), "Smoking and Lung Cancer: Recent Evidence and a Discussion of Some Questions," *Journal of the National Cancer Institute*, 22, 173–203.

Cox, D. R., and Oakes, D. (1984), *Analysis of Survival Data*, London: Chapman and Hall.

DeGroot, M. H., Fienberg, S. E., and Kadane, J. P. (eds.) (1986), *Statistics and the Law*, New York: John Wiley.

Dempster, A. P. (1988), "Employment Discrimination and Statistical Science," *Statistical Science*, 3, 149–195.

Egesdal, S. M. (1986), "The *Frye* Doctrine and Relevance Approach Controversy: An Empirical Evaluation." *Georgetown Law Journal*, 74, 1769–1790.

Eggleston, R. (1983), *Evidence, Proof and Probability* (2nd ed.), London: Weidenfeld and Nicolson.

Faigman, P. L., and Baglioni, A. J., Jr. (1988), "Bayes Theorem in the Trial Process," *Law and Human Behavior*, 12, 1–17.

Fienberg, S. E. (ed.) (1989), *The Evolving Role of Statistical Assessments as Evidence in the Courts*, New York: Springer-Verlag.

Fienberg, S. E., and Straf, M. L. (1982), "Statistical Assessments as Evidence," *Journal of the Royal Statistics Society*, Ser. A, 145, 410–421.

Finkelstein, M. (1978), *Quantitative Methods in Law*, New York: The Free Press.

Finkelstein, M. O., and Levin, B. (1990), *Statistics for Lawyers*, New York: Springer-Verlag.

Fleiss, J. (1981), *Statistical Methods for Rates and Proportions* (2nd ed.), New York: John Wiley.

Freedman, D. A. (1985), "The Mean Versus the Median: A Case Study in 4-R Act Litigation," *Journal of Business & Economic Statistics*, 3, 1–13.

Gastwirth, J. L. (1981), "Estimating the Demographic Mix of the Available Labor Force," *Monthly Labor Review*, 104, 50–57.

——— (1982), "Statistical Properties of a Measure of Tax Assessment Uniformity," *Journal of Statistical Planning and Inference*, 6, 1–12.

——— (1984), "Statistical Methods for Analyzing Claims of Employment Discrimination," *Industrial and Labor Relations Review*, 38, 75–86.

——— (1988), *Statistical Reasoning in Law and Public Policy*, San Diego: Academic Press.

——— (1989), "The Importance of Timing Considerations in the Analysis of Data in EEO Litigation," in *Proceedings of the 41st. Annual Meeting of the Industrial Relations Research Association*, pp. 313–319.

Gastwirth, J. L., and Wang, J. L. (1987), "Nonparametric Tests in Small Unbalanced Samples: Application in Employment Discrimination Cases," *Canadian Journal of Statistics*, 15, 339–348.

Goldstein, R. (1989), "Power and Sample Size via MS/PC-DOS Computers," *The American Statistician*, 43, 253–260.

Hoffman, C. C., and Quade, D. (1983), "Regression and Discrimination: A Lack of Fit," *Sociological Methods and Research*, 11, 407–442.

Hollander, M., and Proschan, F. (1979), *The Statistical Exorcist: Dispelling Statistical Anxiety*, New York, Marcel Dekker.

Juriansz, R. G. (1987), "Recent Developments in Canadian Law: Anti-Discrimination Law Part II," *Ottawa Law Review*, 19, 667–721.

Kaye, D. H. (1987a), "Apples and Oranges: Confidence Coefficients and the Burden of Persuasion," *Cornell Law Review*, 73, 54–77.

——— (1987b), "The Admissibility of Probability Evidence in Criminal Trials-Part II," *Jurimetrics Journal*, 27, 160–172.

Kruskal, W., and Majors, R. (1989), "Concepts of Relative Importance in Recent Scientific Literature," *The American Statistician*, 43, 2–6.

Kuhn, T. S. (1970), *The Structure of Scientific Revolutions*, Chicago: University of Chicago Press.

Lagakos, S. W., Wessen, B. J., and Zelen, M. (1986), "An Analysis of Contaminated Well Water and Health Effects in Woburn, Massachusetts," *Journal of the American Statistical Society*, 81, 583–596.

Lagakos, S. W., and Mosteller, F. M. (1986), "Assigned Shares in Compensation for Radiation-Related Cancers," *Risk Analysis*, 6, 345–357.

Lancaster, H. O. (1949), "The Combination of Probabilities Arising From Data in Discrete Distributions," *Biometrika*, 36, 370–382.

Louv, W. C., and Littell, R. C. (1986), "Combining One-Sided Binomial Tests," *Journal of the American Statistical Association*, 81, 550–554.

McConahay, J., Mullin, C., and Frederick, J. L. (1977), "The Uses of Social Science in Trials With Political and Racial Overtones: The Trial of Joan Little," *Law and Contemporary Problems*, 41, 205–229.

Meier, P. (1986), "Damned Liars and Expert Witnesses," *Journal of the American Statistical Association*, 81, 269–276.

Posner, R. A. (1990), *The Problem of Jurisprudence*, Cambridge, MA: Harvard University Press.

Rosenbaum, P. R. (1987), "The Role of a Second Control Group," *Statistical Science*, 2, 292–316.

Samuel, W. M. (1988), "The Use of Age Classes of Winter Ticks on Moose to Determine the Time of Death," *Canadian Society of Forensic Science Journal*, 21, 54–59.

Savitz, P. A. (1988), "Human Studies of Human Health Hazards: Comparison of Epidemiology and Toxicology," *Statistical Science*, 3, 306–313.

Schafer, D. W. (1987), "Measurement-Error Diagnostics and the Sex Discrimination Problem," *Journal of Business & Economic Statistics*, 5, 529–538.

Solomon, H. (1982). "Measurement and Burden of Evidence," in *Some Recent Advance in Statistics*, ed. J. Tiago de Oliviera, and B. Epstein, New York: Academic Press.

Thompson, W. C., and Schumann, E. L. (1987), "Interpretation of Statistical Evidence in Criminal Trials," *Law and Human Behavior*, 11, 167–187.

Tillers, P. A., and Green, E. (eds.) (1988), *Probability and Inference in the Law of Evidence: The Uses and Limits of Bayesianism*, Dordrecht: Kluwer Academic Publishers.

Weinstein, J. B. (1988), "Litigation and Statistics," *Statistical Science*, 3, 286–297.

CASES CITED

Adams v. Fuqua Industries (1987), 820 F.2d 271 (8th Cir.).

Allen v. United States (1984), 588 F. Supp. 247 (D. Utah) *reversed on legal grounds* 816 F.2d 1417 (10th Cir. 1987).

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

Brock v. Merrill Dow (1989), 874 F.2d 307 (5th Cir.).

Brooks Shoe Company v. Suave Shoe Corporation (1981), 535 F. Supp. 75 (S.D. Fla.).

Capaci v. Katz and Besthoff (1981), 525 F. Supp. 317 (E.d. La.).

Capaci v. Katz and Besthoff (1983), 711 F.2d. 647 (5th Cir.).

Castenada v. Partida (1977), 430 U.S. 482.

Connecticut v. Teal (1982), 457 U.S. 440.

Cooper v. The University of Texas at Dallas (1979), 482 F. Supp. 187 (N. D. Tex.).

Cox v. Conrail (1987), 47 FEP Cases 680 (D.D.C.).

EEOC v. Shelby County Government (1988), 48 FEP Cases 757 (W.D. Tenn.).

Ferebee v. Chevron Chemical Company (1984), 734 F.2d 1259 (D.C. Cir.).

Frazier v. Conrail (1988), 47 FEP Cases 720 (D.C. Cir.).

Frye v. United States (1923), 293 F. 1013 (D.C. Cir.).

Gay v. Waiters (1982), 694 F.2d. 531 (9th Cir.).

Griggs v. Duke Power (1971), 401 U.S. 424.

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

Hopkins v. Price Waterhouse (1989), 109 5 Ct. 1775.

Hopkins v. Price Waterhouse (1990), 52 FEP Cases 1274 (D.D.C.).

Irwin v. Dowd (1961), 366 U.S. 717.

Jackson v. Firestone Tire and Rubber Company (1986), 788 F.2d (5th Cir. 1986).

O'Keefe v. Smith Associates (1965), 380 U.S. 359.

Sulesky v. United States (1980), 545 F. Supp. 426 (S.D.W.Va.).

Tagatz v. Marquette Univ. (1988), 50 FEP Cases 99 (7th Cir.).

Texas Department of Community Affairs v. Burdine (1981), 101 S. Ct. 1089.

Twin Manors Condominium Association v. Rosewell (1988), 529 N.E. 1104, *cert. den.* 57 *U.S. Law Week* 3859.

United States v. Fatico (1978), 458 F. Supp. 338 (S.D. N.Y.).

United States v. 43 1/2 Gross of Rubber Prophylactics (1946), 65 F. Supp. 534 (D. Minn.).

United States v. Haldeman (1976), 559 F.2d. 31 (D.C. Cir.).

Vuyanich v. Republic National Bank (1980), 505 F. Supp. 224 (N.D. Texas).

Wards Cove Packing Company et al. v. Frank Antonio et al. (1989), 109 S. Ct. 2115.

Warner-Lambert v. Hechler (1986), Food and Drug Decisions 38, 346 (3rd. Cir.).

References

Comparability of Randomised Groups
Douglas G. Altman
The Statistician, Vol. 34, No. 1, Statistics in Health. (1985), pp. 125-136.
Stable URL:
http://links.jstor.org/sici?sici=0039-0526%281985%2934%3A1%3C125%3ACORG%3E2.0.CO%3B2-5

Evolving Legal Standards for the Admissibility of Scientific Evidence
Bert Black
Science, New Series, Vol. 239, No. 4847. (Mar. 25, 1988), pp. 1508-1512.
Stable URL:
http://links.jstor.org/sici?sici=0036-8075%2819880325%293%3A239%3A4847%3C1508%3AELSFTA%3E2.0.CO%3B2-Z

Some Methods for Strengthening the Common #2 Tests
William G. Cochran
Biometrics, Vol. 10, No. 4. (Dec., 1954), pp. 417-451.
Stable URL:
http://links.jstor.org/sici?sici=0006-341X%28195412%2910%3A4%3C417%3ASMFSTC%3E2.0.CO%3B2-1

Employment Discrimination and Statistical Science
Arthur P. Dempster
Statistical Science, Vol. 3, No. 2. (May, 1988), pp. 149-161.
Stable URL:
http://links.jstor.org/sici?sici=0883-4237%28198805%293%3A2%3C149%3AEDASS%3E2.0.CO%3B2-1

Bayes' Theorem in the Trial Process: Instructing Jurors on the Value of Statistical Evidence
David L. Faigman; A. J. Baglioni, Jr.
Law and Human Behavior, Vol. 12, No. 1. (Mar., 1988), pp. 1-17.
Stable URL:
http://links.jstor.org/sici?sici=0147-7307%28198803%2912%3A1%3C1%3ABTITTP%3E2.0.CO%3B2-V

The Mean versus the Median: A Case Study in 4-R Act Litigation
D. A. Freedman
Journal of Business & Economic Statistics, Vol. 3, No. 1. (Jan., 1985), pp. 1-13.
Stable URL:
http://links.jstor.org/sici?sici=0735-0015%28198501%293%3A1%3C1%3ATMVTMA%3E2.0.CO%3B2-F

Statistical Methods for Analyzing Claims of Employment Discrimination
Joseph L. Gastwirth
Industrial and Labor Relations Review, Vol. 38, No. 1. (Oct., 1984), pp. 75-86.
Stable URL:
http://links.jstor.org/sici?sici=0019-7939%28198410%2938%3A1%3C75%3ASMFACO%3E2.0.CO%3B2-%23

Power and Sample Size via MS/PC-DOS Computers
Richard Goldstein
The American Statistician, Vol. 43, No. 4. (Nov., 1989), pp. 253-260.
Stable URL:
http://links.jstor.org/sici?sici=0003-1305%28198911%2943%3A4%3C253%3APASSVM%3E2.0.CO%3B2-9

Concepts of Relative Importance in Recent Scientific Literature
William Kruskal; Ruth Majors
The American Statistician, Vol. 43, No. 1. (Feb., 1989), pp. 2-6.
Stable URL:
http://links.jstor.org/sici?sici=0003-1305%28198902%2943%3A1%3C2%3ACORIIR%3E2.0.CO%3B2-S

The Combination of Probabilities Arising from Data in Discrete Distributions
H. O. Lancaster
Biometrika, Vol. 36, No. 3/4. (Dec., 1949), pp. 370-382.
Stable URL:
http://links.jstor.org/sici?sici=0006-3444%28194912%2936%3A3%2F4%3C370%3ATCOPAF%3E2.0.CO%3B2-O

Combining One-Sided Binomial Tests

William C. Louv; Ramon C. Littell

Journal of the American Statistical Association, Vol. 81, No. 394. (Jun., 1986), pp. 550-554.

Stable URL:

http://links.jstor.org/sici?sici=0162-1459%28198606%2981%3A394%3C550%3ACOBT%3E2.0.CO%3B2-K

The Uses of Social Science in Trials with Political and Racial Overtones: The Trial of Joan Little

John B. McConahay; Courtney J. Mullin; Jeffrey Frederick

Law and Contemporary Problems, Vol. 41, No. 1, Criminal Process in the Seventies. (Winter, 1977), pp. 205-229.

Stable URL:

http://links.jstor.org/sici?sici=0023-9186%28197724%2941%3A1%3C205%3ATUOSSI%3E2.0.CO%3B2-4

Damned Liars and Expert Witnesses

Paul Meier

Journal of the American Statistical Association, Vol. 81, No. 394. (Jun., 1986), pp. 269-276.

Stable URL:

http://links.jstor.org/sici?sici=0162-1459%28198606%2981%3A394%3C269%3ADLAEW%3E2.0.CO%3B2-1

The Role of a Second Control Group in an Observational Study

Paul R. Rosenbaum

Statistical Science, Vol. 2, No. 3. (Aug., 1987), pp. 292-306.

Stable URL:

http://links.jstor.org/sici?sici=0883-4237%28198708%292%3A3%3C292%3ATROASC%3E2.0.CO%3B2-H

Human Studies of Human Health Hazards: Comparison of Epidemiology and Toxicology

David A. Savitz

Statistical Science, Vol. 3, No. 3. (Aug., 1988), pp. 306-313.

Stable URL:

http://links.jstor.org/sici?sici=0883-4237%28198808%293%3A3%3C306%3AHSOHHH%3E2.0.CO%3B2-L

Measurement-Error Diagnostics and the Sex Discrimination Problem

Daniel W. Schafer

Journal of Business & Economic Statistics, Vol. 5, No. 4. (Oct., 1987), pp. 529-537.

Stable URL:

http://links.jstor.org/sici?sici=0735-0015%28198710%295%3A4%3C529%3AMDATSD%3E2.0.CO%3B2-R

LINKED CITATIONS

- Page 4 of 4 -

Interpretation of Statistical Evidence in Criminal Trials: The Prosecutor's Fallacy and the Defense Attorney's Fallacy

William C. Thompson; Edward L. Schumann

Law and Human Behavior, Vol. 11, No. 3. (Sep., 1987), pp. 167-187.

Stable URL:

http://links.jstor.org/sici?sici=0147-7307%28198709%2911%3A3%3C167%3AIOSEIC%3E2.0.CO%3B2-9

Litigation and Statistics

Jack B. Weinstein

Statistical Science, Vol. 3, No. 3. (Aug., 1988), pp. 286-297.

Stable URL:

http://links.jstor.org/sici?sici=0883-4237%28198808%293%3A3%3C286%3ALAS%3E2.0.CO%3B2-V