SOME STATISTICAL CONCEPTS AND IDEAS THAT I WISH JUDGES AND LAWYERS KNEW

Joseph L. Gastwirth, Department of Statistics, George Washington University, USA

The importance of statistical issues in the legal decision making process is illustrated by examining actual decisions.

INTRODUCTION

Although probabilistic and statistical concepts and terminology, such as expectation damages, have long been used in the legal area only in the last twenty-five years or so has statistical evidence and its concomitant analytic tools been routinely introduced in a wide variety of cases. As the Federal Judicial Center's (1994) guide to scientific evidence includes several chapters on statistics, including surveys, survival and case-control studies, regression analysis and the basic concepts of estimation and hypothesis testing, and discusses their evidentiary role, we focus on issues arising in the statistical analysis of data from actual cases.

THE IMPORTANCE OF TRANSLATING THE LEGAL ISSUE INTO A STATISTICAL PROBLEM: DEFINING THE PARAMETER OF INTEREST

Disparate impact cases examine whether a minority group has a significantly lower probability of satisfying a particular employment practice, e.g., passing a test or possessing a pre-set level of education. If so, the practice must be shown to be job-related, i.e., predictive of successful job performance. Meier et al. (1984) justify the use of the four-fifths rule, i.e., the ratio of minority to majority pass rates should be less than .8 before a practice is deemed to have a disparate impact. Blind adherence to it, however, in cases concerning layoffs lead to anomalous decisions. Gastwirth and Greenhouse (1995) illustrate this point on data from *Council 31 v. Ward*, 60 FEP Cases 275 (7th Cir. 1992).

The government decided to concentrate the closing of its offices in the Chicago area. Thus, 130 black employees from a pool of 1512 were laid off while 87 out of 2900 white employees were laid off. In the case 8.6 per cent of the blacks were terminated while only 3 per cent of the whites were: the plaintiffs argued that the ratio of the termination rates 3/8.6=0.349 easily met the 'four-fifths' rule. The defendant argued that retention rates should be used: as 91.4 per cent of black employees kept their jobs and 97 per cent of whites, the 'selection ratio' of 0.942 exceeded the 'four-fifths' rule. The lower

court accepted the defendant's claim but the appellate opinion, remanded the case for reconsideration. Notice that if *all* 217 employees laid off were black, 85.36 per cent of the blacks would be retained, so that if the court accepted the defendant's argument, it would find no discrimination. The *p*-value of Fisher's exact test is essentially zero and the 95% CI, obtained from STATXACT, is (2.815, 4.068).

The 'four-fifths' rule arose in the context of evaluating the impact of a pre-employment test or criterion. If 98 per cent (96 per cent) white (minority) applicants pass the test or meet the criterion, even though members of the minority group have about half the odds of passing as whites, a court might well find the net effect legally minimal. Thus, whether the selection ratio or the odds ratio is appropriate depends on the legal issue involved; the odds ratio is more suitable for termination cases while the selection ratio seems preferable when comparing hiring or promotion rates. The ongoing case of *Bew v. City of Chicago* 75 FEP Cases (DC N.Ill 1997) illustrates that it may not be possible to set a pre-determined value of a measure or a pre-set level of significance that is applicable in all situations. After completing a training class in order to become a member of the police force one had to pass a state exam. The pass rates of both blacks and whites were very high and the black rate was 98.16% that of the whites. As the number of test takers was large, the standard test yielded a difference of 5 standard deviations, clearly a significant one. The defendant moved for summary judgment on the basis that the difference was not legally meaningful in light of the 'four-fifths' rule while the plaintiffs emphasized the statistically significant difference between the pass rates. The judge decided that a full trial should be held.

In the equal employment context a selection ratio of 0.98, as in *Bew*, would *not* usually be considered meaningful, especially as so many minority officers passed the exam. The same data, however, might be rather strong evidence in a negligence case. Suppose the data concerned two screening tests for HIV infection. Test A applied to AIDS patients had the data for whites (pass means infection detected) and Test B had the data for blacks. Clearly, someone receiving a transfusion would choose Test A. Suppose a hospital knew this but had over 10000 test B kits and deliberately decided to continue to use them up before telling patients about test A. A patient who becomes infected with HIV after a routine operation might well have a reasonable claim of negligence against the hospital for not using Test A or at least informing patients of its availability.

STATISTICAL SIGNIFICANCE AND POWER AND THEIR RELATIONSHIP TO

SAMPLE SIZE

In the U.S. courts have become comfortable with the calculation of statistical significance as laid out by the Supreme Court in the *Castenada v. Partida* jury discrimination case. The meaning of p-value or the significance level of a test, however, has been misinterpreted as meaning the probability that the null hypothesis is true. It should be emphasized that it is the probability that a result at least as far from expected, under the null hypothesis of fairness, as the observed one would occur. We cannot quantify the probability that the null hypothesis is true without assuming a prior probability for it.

The concept of power or its complement the Type II error is rarely considered. In the *Capaci* case discussed in Finkelstein and Levin (1990) a court accepted a test that had *zero* power to detect a difference in time to promotion between males and females even though a more powerful and appropriate test found a difference of about 10 years significant. Before describing the data I should mention that I was the plaintiffs' expert in the case and used the Wilcoxon test to demonstrate that the difference in time to promotion of the two women and 24 men was significant as the females had the 24th and 25th longest waiting times (one-sided p-value = 0.0123). The defendant's expert used the median test and concluded that the difference in the two distributions of time to promotion was not significant. With only two females in the data set the median test could *never* find a difference significant. This might have been brought out to the trial judge and made clear in the record by the following sequence of questions:

1. The difference between the average times or the median times it took women and men to be promoted was about eight years, would the median test you used have found a difference of 10 years significant? Ans. No, it would not.
2. Would the median test have found a difference of 20 years between the two average times significant? Ans. No.
3. Repeat for 50 years, 100 years, 1000 years, one million years. The answer would remain no, it would not.
4. Then, plaintiffs' lawyer turns to the expert and points to the named plaintiff and asks "Just how long must my client and other women work without a promotion before your median test would determine that there was a significant difference between the promotion times of males and females?"

Further discussion of the role of power, the need to carefully examine "explanations" of significant results, timing considerations and the importance of preserving information relevant to damage calculations appears in Gastwirth (1998). We next turn to an issue from a recent Supreme Court decision.

ISSUES IN COMBINING STRATIFIED DATA OR RESULTS FROM SEVERAL STUDIES

In both discrimination and product liability cases one should compare the minority and majority job applicants or the exposed and unexposed persons after adjusting for relevant covariates, e.g., education in the first case and duration of exposure to the agent as well as exposure to other toxic chemicals in the second. One approach is to stratify the data into comparable subgroups. While combination methods such as the Mantel-Haenszel test and related estimate of the common odds ratio and Fisher's summary chi-square test have been accepted by courts in discrimination cases, the meta-analysis of several related epidemiologic studies has not been as well understood. Indeed, in the recent *General Electric v. Joiner*, 66LW 4036 (1997) case only Justice Stevens understood that to do a proper meta-analysis all studies should be available. Indeed, the statistical literature on the subject not only emphasizes this point but has been concerned with the potential effect on a meta-analysis when studies not showing a significant result may not be published and hence are missing from the data base (Iyengar and Greenhouse, 1987; Berlin, Begg and Louis, 1989).

Before discussing methodology, the need to combine studies should be explained. Unlike designed experiments or surveys where the sample size can be determined by the investigator, in an epidemiologic study examining whether workers exposed to a chemical on the job have a higher incidence of a disease the number of workers employed in a specific position depends on the production process and on the economic demand for the product. Thus, the available sample may be quite small. Consequently, the power of a single study to detect a relative risk of 2, say, may not even reach 50% so it is essential that the results of several similar studies are considered as a whole. The same issue arises in multi-center clinical trials evaluating a new treatment of a rare disease. No single center will have a sufficient number of patients eligible for the study in order for the study to have adequate power to detect a meaningful improvement in survival under the new drug.

In the *Joiner* case, which concerned whether PCBs caused the plaintiff's cancer, the District Court did not admit the plaintiff's experts' testimony. One expert claimed that Joiner's cancer was more likely due to PCB exposure than to his smoking while the second testified that materials with which Joiner worked caused or significantly contributed to his lung cancer. Apparently, there were at least 13 different studies relied on by the plaintiff's experts. Justice Stevens points out that only one of them was in the record and only six were discussed by the District court. The trial judge examined these studies individually and concluded that no single study showed a statistically significant increased risk of lung cancer due to exposure to PCBs. The Appeals Court decided that a "weight of the evidence" or meta-analytic approach was scientifically acceptable, however, the Supreme Court majority followed the trial court's approach when it discussed four studies. After questioning whether a fair appraisal of the scientific methodology used by plaintiffs' experts could be made by looking at half of the studies, Justice Stevens notes that while he agrees with both the District Court's and the majority that each of the studies by itself was unpersuasive the key question was left unanswered: Why were the opinions of qualified experts relying on a scientifically acceptable method, meta-analysis or combining the results in the relevant scientific literature, inadmissible?

Three statistical calculations may shed light on why all the studies should be examined and their results summarized in an overall result. We will use a test, however, similar methods exist for developing an estimate and associated CI for a common overall measure of risk such as the odds ratio or relative risk. Suppose one had available three independent, small, studies, each of which was not significant at the standard 0.05 level but had a p-value of 0.10. Fisher's summary chi-square statistic on 6df equals 13.816 , which is statistically significant at the 0.05 level. Collectively, the studies do indicate a significant association between exposure and disease. On the other hand, suppose there had been one study which did reach statistical significance with a p-value of 0.04 but the other two studies had p-values of 0.45 and 0.40, say, respectively. Now, the summary chi-square test equals 9.87 , which is not statistically significant (overall p-value slightly exceeds 0.10). Thus, relying on only the one "significant" study could lead to an erroneous inference. When three studies are carried out, each of which has probability 0.05 of yielding a "significant" result the probability that, by chance, at least one will be significant is 0.143, noticeably larger than the usual 0.05 level. Finally, if one required all three studies to be significant, one would be using a significance level of 0.05 or

0.000125, which is absurdly low and renders it virtually impossible to detect a meaningful difference.

REFERENCES

Berlin, J. A., Begg, C. B. and Louis, T. A. (1989). An assessment of publication bias using a sample of published clinical trials. *Journal of American Statistical Association, 84,* 381-392.

Federal Judicial Center (1994). *Reference Manual on Scientific Evidence.* Washington, DC: US Government Printing Office.

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

Gastwirth, J. L. (1998). *Statistical concepts and ideas I wished judges and lawyers knew.* Technical Report, The George Washington University.

Gastwirth, J.L. and Greenhouse, S. W. (1995). Biostatistical concepts and methods in the legal setting. *Statistics in Medicine 14*, 1641-1653.

Iyengar, S. and Greenhouse, J. B. (1988). Selection models and the file drawer problem (with discussion). *Statistical Sciences, 3,* 109-135.

Meier, P., Sacks, J. and Zabell, S. L. (1984). What happened in Hazelwood? Statistics, employment discrimination and the law. *American Bar Foundation Research Journal 1*, 139-186.