i-lex. Scienze Giuridiche, Scienze Cognitive e Intelligenza artificiale
Rivista quadrimestrale on-line: www.i-lex.it
Novembre 2012, numero 17

THE NEED FOR COURTS AND POLICY MAKERS TO CAREFULLY EVALUATE CRITICISMS OF STATISTICAL EVIDENCE

Joseph L. Gastwirth [*]

Abstract. In virtually every reasonably large scale statistical study inevitably some errors in the data or modest deviations from the assumptions made in the analysis occur. Thus, it is easy to raise a question about the conclusions. Sometimes courts and policy makers postpone decisions, e.g. ask for a further epidemiologic study or give an otherwise sound statistical analysis much less weight than it deserves. It is preferable for decision-makers to require critics to demonstrate that the potential flaws they mention are sufficiently severe that they would affect the main inferences drawn from the study. A statistical approach to assessing the potential impact of an omitted characteristic on the conclusions of a study is illustrated and supports legal decisions that did not simply accept a suggested criticism. An example where a court accepted a suggested "explanation" of a statistically significant disparity, which mathematically could not reduce the disparity to non-significance is given.

Key-words. Equal employment, omitted variable, scientific evidence, sensitivity analysis, statistics in law

1. Introduction

Important policy and legal decisions are based in part on statistical studies, which are submitted as evidence in cases or in the record of an administrative agency's proceedings. The scientific reliability of these studies is often questioned by the opposing party in a legal case or a business-related organization whose members could be affected by a new law or regulation. Some of the areas where statistical analyses are considered are: environmental[1], product liability[2], securities law[3],

[*] Department of Statistics, George Washington University.

Acknowledgements: It is a pleasure to thank Ms.Wenjing Xu for several helpful discussions and Dr. Alessandro Serpe for his encouraging interest in this area of legal argumentation.

Suggested citation: Joseph L. Gastwirth, The Need for Courts and Policy Makers to Carefully Evaluate Criticisms of Statistical Evidence in: i-lex, 17, 2012, pp.235-248 (www.i-lex.it)

Invited paper

trademark infringement[4], international trade[5] and discrimination against minorities[6], In the 1980's several legal decisions in employment

[1] See: *Coalition for Responsible Regulation v. EPA*, D.C. Cir. No. 09-1322 (2012) and *Massachusetts v. EPA*, 549 U.S. 497(2007) (concerning the regulation of greenhouse gases by the EPA).

[2] *Borel v. Fiberboard Paper Products, Corp*., 493 F.2d 1076 (5th Cir, 1973) is a classic opinion dealing with asbestos related diseases. For recent developments, see: DOMINICI, F., KRAMER, S. and ZAMBELLI-WIENER, A. (2008), *The role of epidemiology in the law: a toxic tort litigation case*. GASTWIRTH, J.L. , *Law, Probability and Risk*, 7, 2003, pp. 15-34; *The need for careful evaluation of epidemiologic evidence in product liability cases*, in: *Law, Probability and Risk*, 2, pp. 151-189 and MILLER, C. (2012), *Epidemiology in the courtroom: mixed messages from recent British experience* in: *Law, Probability and Risk*, 11, p.p. 85-100 and the *Reference Manual on Scientific Evidence* (2nd ed. 2000) issued by the Federal Judicial Center. The Chapters on Statistics, Regression analysis, Epidemiology and Surveys are most relevant to the topics discussed here.

[3] *Pearl v. Geriatric & Medical Centers, Inc*. 1996 U.S. Dist. LEXIS 1559 (E.D, Pa. 1996) (regression analysis used to estimate damages in a fraud on the market case).

[4] Zippo Manufacturing Co. V. Rogers Imports, Inc. 216 F. Supp. 670 (S.D. N.Y. 1963) is a classic case relying on a survey to demonstrate consumer confusion. See Shari S. Diamond, *Reference Guide on Survey Research* in Federal Judicial Center, *Reference Manual on Scientific Evidence* 229 (2nd ed. 2000) and GASTWIRTH, J.L., *Issues Arising in Using Samples as Evidence in Trademark Cases*, in: *Journal of Econometrics*, 113, 2003, pp. 69-82 for a discussion of more recent cases.

[5] The methodology for calculating fair market values to determine the appropriate costs in anti-dumping cases from producer's in non-market economies was at issue in *Dorbest Ltd. V. United States*, 604 F.3d 1363 (Fed. Cir. 2010) and *Diamond Sawblades Mfrs. Coalition v. United States*, 626 F. 3d 1374 (Fed. Cir. 2010).

[6] The U.S. Supreme Court adopted statistical hypothesis testing in *Castaneda v. Partida*, 430 U.S. 482 (1977), an equal protection case concerning jury discrimination against Mexican-Americans. The procedure compares the proportion of venire members who are minority with the corresponding proportion of the jury eligible population in the jurisdiction. For a description of the methodology see Finkelstein, M 0. (1966) The application of statistical decision theory to the jury selection cases. *Harvard Law Review*, 80, pp. 338-376 and KAYE, D. H. (1985) Statistical analysis in jury discrimination cases. *Jurimetrics*, 25, pp. 276-289. A related issue concerns the representativeness of juries, which arises in cases brought under the Sixth Amendment and laws

discrimination cases noted that a critic needs to do more than raise a question about a potential flaw in a statistical study, they should demonstrate that correcting for it could change the conclusions drawn from the study[7]. This paper reviews one approach for assessing whether a variable or characteristic that was not considered in a study could alter the ultimate inference. The data from two cases are reanalyzed and it will be seen that the offered explanation was insufficient to explain the disparity in both. In one case, our analysis confirms the legal decision, which did not accept the proposed explanation. In the second, the court accepted a highly implausible explanation. Tin Section 3 the data from a recent equal pay case in which the 7th Circuit Court of Appeals reversed a lower court's decision to accept an employer's suggestion that education and prior experience justified a substantial disparity *without* submitting an analysis including those variables. A regression analysis of the available information, which was *not* submitted in evidence, will be seen to support the appellate decision.

2. Illustrative uses of Cornfield's approach to assess the potential impact of an omitted variable on data from real cases

In epidemiology, the ratio (R) of the incidence of a disease in an exposed group to that of an unexposed group is of concern. After a number of studies indicated that smoking substantially increases the risk of lung cancer, the industry suggested many other potential factors that might explain the five to ten fold increased relative risk that smokers

implementing it. The main case stating the criteria used to evaluate these claims is Duren v. Missouri, 439 U.S. 357 (1979). The Court has not yet specified its preferred statistical method to examine data in these cases, see *Berghuis v. Smith*, 130 S. Ct. 1382 (2010). The statistical data submitted by the defendant's expert is reanalyzed by GASTWIRTH, J.L. and PAN, Q. (2011). Statistical measures and methods for assessing the representativeness of juries: a reanalysis of the data in *Berghuis v. Smith*. *Law, Probability and Risk*, 10, pp. 17-57, to account for the fact that the racial composition of the jury needed to be estimated because the main source of jurors in Michigan is the list of individuals who have a driver's license or identification card from the Department of Motor Vehicles. Michigan and Illinois are the only two states in the United States that do not include racial identification on driver's licenses.

[7] See *Capaci v. Katz and Besthoff*, 711 F. 2d 647 (5th Cir. 1983), *Palmer v. Shultz*, 815 F. 2d, 84, 101 (D.C. Cir., 1987) *Allen v. Seidman*, 881 F.2d 375, 379-80 (7th Cir. 1989).

had relative to non-smokers in most studies. Cornfield set out criteria that such an omitted variable or factor needed to satisfy in order for it to "explain" the observed increased risk[8]. For examining data from equal employment cases, one denotes the success (e.g., pass, promotion) rate of the minority (majority) group by $p_1(p_2)$ and their ratio, $R=p_2/p_1$.

Cornfield's Lemma: In order for a factor U to explain a disparity between two rates, the factor must multiply one's chance of success by at least R *and* the prevalence of U in the majority group must be at least R times its prevalence in the minority group.

A strengthening of the second condition occurs when the prevalence (f_1) of the characteristic U is known. Then the prevalence (f_2) must satisfy

$$(1) \qquad f_2 \geq Rf_1 + \frac{R-1}{R_U-1}$$

where R_U is the strength of the association between having U and success, i.e. having U multiplies one's probability of success by R_U.

To illustrate the usefulness of the inequality, consider the data from *Allen v*. *Seidman*[9]. The plaintiffs were Black bank examiners who alleged that the promotion examination given by the Federal Deposit Insurance Corp had a disparate impact on them. Under the disparate impact approach, the plaintiff needs to show a significant difference in the pass rates of minority and majority applicants. Typically this is done by showing that the difference is statistically significant at the two or three standard deviation level, which approximately corresponds to significance at the .05 or .01 level. Once the plaintiff satisfies this requirement the employer needs to show that the exam is job-related.[10]

---

[8] S.W. GREENHOUSE (1982). *Jerome Cornfield's contributions to epidemiology*., in: *Biometrics*, 38, Suppl. pp. 33-45.

[9] 881 F.2d 375 (7th Cir. 1989).

[10] There is a large literature concerning disparate impact cases. Three recent articles provide an overview and many references are: Ngov, E.N. (2011). When "The Evil Day" Comes, Will Title VII's Disparate Impact

Provision Be Narrowly Tailored to Survive an Equal Protection Challenge*?*, 60, *American. Univ. L. Review* 535-588. Sullivan, C. A.(2010) *Ricci v. DeStefano*: End of the line or just another turn in the disparate impact road? *Northwestern Law Review*, 104, 411- 426 and Seiner, J.A. and Gutman, B.N. (2010). Does Ricci herald a new disparate impact?, 90, *Boston Univ. Law Review*, 2181-2213. An alternative analysis of the pass rate data submitted in the Ricci case is given by Gastwirth, J.L. and, Miao, W. (2009). Formal statistical analysis of the data in

All the candidates had worked for between 5 and 15 years and had obtained a recommendation letter from their regional director. Only 14 of 36 Black applicants passed the exam, while 329 Whites out of 391 passed, Thus, 38.89% (84.14%) of Blacks (Whites) passed, so R= 2.16 and the difference is highly significant (p-value < 0001). Suppose the defendant suggested that a factor, e.g. having a Master's degree in Business increased one's probability of passing a promotion exam by a factor of 3. Then if only 10% of the Blacks had such a degree, the second condition implies that the fraction of Whites possessing the degree would need to be at least

(2.16)(.10) + (1.16)/2 = .796.

One can account for sampling error by using the lower end of a 95% confidence interval for R[11]. This analysis supports the decision that found that the data supported the claim that the test had a disparate impact on minorities, so the defendant would need to demonstrate that it was job-related, i.e. predictive of job success. The decision noted that the FDIC had not submitted an analysis, e.g. a logistic regression that included another job-related predictor, such as higher relevant education, that explained the disparity.

By clearly stating the criteria another job-related characteristic needs to satisfy in order for it to explain a statistically significant finding, Cornfield's approach enables courts and policy makers to evaluate the plausibility of a proposed "explanation". In some situations, a modest imbalance in the prevalence of a factor with a modest relationship to the response of interest (success in the employment discrimination context) will suffice. This probably occurred in the early studies that showed a

---

disparate impact cases provides sounder inferences than the U.S. government's 'four-fifths' rule: An examination of the record in *Ricci v. DeStefano*. Law, Probability and Risk, 8, 171-191.The 'four-fifths' rule is a government guideline stating that whenever the ratio of the minority pass rate to the majority pass rate is less than four-fifths or 80%, the exam or employment practice has a disparate impact. In small to moderate sample sizes, the sampling fluctuations in the pass rates of a "fair" test can cause it to fail this criterion more than 20% of the time.

[11] See: GASTWIRTH, J.L. (1992) Methods for assessing the sensitivity of statistical comparisons used in Title VII cases to omitted variables , 33, pp. 19-34 where accounting for sampling error reduces the required prevalence of 79.6% to 70%, which is not plausible. A comprehensive treatment of the subject is given by P.R. ROSENBAUM (2002) *Observational Studies* (2nd Ed.) NY: Springer.

modest increased risk of birth defects in users of the morning sickness drug, Bendectin[12].

The *Maloley v. National Revenue Service* (1986)[13] case also concerned the disparate impact of an exam. The Canadian Public Service Commission changed the passing score for tax enforcement clerks and on the next exam the pass rate of female employees dropped much more than males. A woman candidate sued under the disparate impact theory, which Canada adopted from the United States. Thus, if the difference in pass rates meets the "two to three" standard deviation criterion, the plaintiff will establish a *prima facie* case.

In *Maloley*, 68 of 251 females (27.1% passed the exam, while 68 of 115 males (59.1%) passed. The difference of 32% is highly significant (p-value of Fisher's exact text is < one in a million).The Revenue Service suggested that the disparity arose because 52% of male applicants had *some* college education but only 27% of the females did. The Service was *not* asked to submit the full data set giving the pass rates for applicants with or without some college by gender.

Let us apply Cornfield's lemma to the data. The omitted variable, U, is having some college education. The observed ratio, R, of majority to minority pass rates is .5913/.2719=2.18. What is the required relationship, R_U of some college education, to passing the exam for this to "explain" the disparity? Here, the prevalence of U in both groups is known. Substituting f_1 = .27, f_2 =.52 and R=2.16 in (1) yields R_U needs to be at least 16.2. Replacing 2.16 by the lower limit (1.69) of a 95% confidence interval for R to account for "*sampling fluctuation*"[14] indicates that R_U needs to be at least 8.19. Is it plausible that just attending college for a couple of years would increase one's probability of passing an exam by a factor of *eight*?

Because the prevalence of U is known, it is possible to find the allocation of exam passers of the both genders by education that is most favorable to the defendant[15], i.e., gives the largest p-value of the

---

[12] M.D. GREEN (1998) *Bendectin and Birth Defects*, Univ. of Pennsylvania Press, Philadelphia.

[13] JURIANSZ, R.G. Jr. (1987). *Recent developments in Canadian Law: Anti-discrimination law: Parts 1&2*, in: *Ottawa lLaw Review*,19, pp. 447, 667.

[14] In this context, one assumes that the applicants were a sample of the eligible employees.

[15] GASTWIRTH, J.L., KRIEGER, A. M. AND ROSENBAUM, P.R. (1994). How a court accepted an impossible explanation, *American Statistician*, 48, 313-316. By "*impossible*" the authors meant that it was impossible for the explanation to

appropriate CMH-statistical test combining the exam results in both tables[16]. The two tables are given in Table 1.

*Table 1: Most Favorable Allocation of the Passing Data by Educational Leve*l

	Some College			*No College*		
	P	*F*	*Total*	*P*	*F*	*Total*
Female	*60*	*3*	*63*	*8*	*180*	*188*
Male	*56*	*4*	*60*	*12*	*43*	*55*
Total	*116*	*7*	*123*	*20*	*223*	*243*

Applying the CMH test to the two 2x2 tables of data in Table 1 yields a value of -3.11 standard deviations (p-value =.003). Thus, the proposed explanation *cannot* reduce the 32% difference in pass rates in the original data sufficiently to meet the most stringent three standard deviation level, much less the frequently adopted two standard deviation level. The sensitivity analysis utilizing the Cornfield approach enables us to evaluate whether a proposed explanation can plausibly reduce a statistical significant and meaningful disparity to a non-significant one.

Comment: This analysis also demonstrates the wisdom of legal decisions that require the party proposing an explanation of a statistical disparity to substantiate it with an analysis incorporating that variable. In the equal employment context, the defendant (employer) should have a record of the factors used in making hiring and promotion decisions, so it is reasonable for courts to expect them to use that information to rebut a statistically significant disparity in minority hires, promotion or pay. The plaintiff, who has access during the discovery phase of litigation, should also be required to utilize all the information

reduce the disparity by an amount sufficient to reduce the statistical significance of the data to less than three standard deviations.

[16] The Cochran-Mantel-Haenszel test is described in: AGRESTI, A. (1990). Categorical Data Analysis, NY: John Wiley at 230-234. Other illustrations of its application to data from legal cases are given in FINKELSTEIN, M.O. and LEVIN, B. J. (2001) Statistics for Lawyers (2d Ed NY: Springer and GASTWIRTH, J.L. (1988) *Statistical Reasoning in Law and Public Policy*, Vol. 1, Academic Press, Boca Raton.

on job-related factors in the analysis or explain why some information was not used[17].

3. An Equal Pay case: *King v. Acosta Sales and Marketing Inc*[18]

Acosta Sales is a food broker representing producers seeking to sell to supermarkets and other bulk purchasers. In 2001, Ms. King was hired as a business manager, i.e. a representative of several producers who are clients of the company. The trial court granted summary judgment to the defendant and that decision was reversed by the appellate decision discussed here. Because motions for summary judgment are often brought at a relatively early stage of legal proceedings, some relevant information is not available; presumably further discovery will be allowed before the full trial is held.

After leaving the company in 2007, Ms. King charged Acosta with sex discrimination in pay, i.e. women doing the same work as men receive lower salaries. If true, this employment practice violates both Title VII of the U.S. Civil Rights Act and the Equal Pay Act, 29 U.S. C. Sec. 206(d). In support of her claim, the plaintiff submitted salary data for each male

[17] In *Boykin v. Georgia-Pacific Corp*. 706 F. 2d 1384 (5th Cir.1983) the plaintiff's expert used all the information, e.g. prior job-related experience, that was asked in the applications form. Minority applicants with experience or without experience had lower higher rates to the better positions than comparable Whites. The defendant criticized the plaintiffs' analysis because it did not standardize the data for additional qualifications. The court did not accept this explanation because most of the jobs were unskilled and were promoted on the basis of the training they received on the job. Apparently, the defendant did not submit a study, e.g. a regression or a stratified analysis, using the CMH test, incorporating these "*additional qualifications*". In some cases plaintiffs have not used information on clearly relevant factors, on which information was available, and courts have correctly not allowed them to establish a *prima facie* case. For example, in *Sheehan v. Daily Racing Form*, 104 F.3d 940 (7th Cir. 1997), the company acquired another firm that was computerized and decided to convert its own publication to a similar computerized system. As a consequence of the acquisition it had a layoff. The plaintiff, who was over 40 claimed that older employees were laid off at a higher rate than younger ones. The court did not credit this finding because it failed to incorporate information on the computer skills of the employees. It observed that younger persons are generally more familiar with computers than older people.

[18] 678 F.3d 470, 114 FEP Cases 897 (7th Cir. 2012).

and female business manager employed during her time at the firm. The data is given in Table 2.

Table 2: Salary Data for Business Managers of Acosta Sales in 2001-2007

StartingYear	*StartingSalary*	*2007 or FinalSalary*	*Gender*
1998	*91,000.08*	*122,004.00*	*0*
2000	*95,000.00*	*101,921.00*	*0*
2004	*85,000.01*	*99,500.11*	*0*
2001	*94,999.99*	*97,635.55*	*0*
2006	*93,000.00*	*93,000.00*	*0*
1998	*69,448.56*	*81,502.73*	*0*
2002	*77,182.51*	*79,881.10*	*0*
1998	*72,799.92*	*79,598.69*	*0*
1998	*63,000.00*	*72,375.05*	*0*
2001	*38,666.64*	*60,399.62*	*1*
2007	*60,000.00*	*60,000.00*	*0*
2005	*40,000.01*	*60,000.00*	*0*
2007	*55,000.00*	*55,000.00*	*0*
2005	*40,000.01*	*52,299.77*	*1*
2001	*40,000.01*	*46,850.23*	*1*
2005	*45,000.00*	*46,350.00*	*1*
2007	*42,500.64*	*42,500.64*	*1*
2007	*40,000.42*	*40,000.42*	*1*
2001	*37,752.00*	*64,000.01*	*1*
2001	*26,624.00*	*38,092.01*	*1*

Note: Female (male) employees are denoted by gender =1(0). The data are given in the opinion, 114 FEP Cases at 899. The author noticed that the starting salaries for the last two employees were less than the salary the received later, so the numbers were switched in Table 2.

The appellate court observed that the difference in salaries between men and women was striking. All of the men were paid more than but one of the women and that woman achieved her $60,000 salary after six years on the job while the men reached that salary level in less time.

The company stated that education and experience account for the higher male salaries. All the men had college degrees, while Ms. King did not (the educational level of the other women was not in the record). The District Court thought it was sufficient for the defendant to articulate education and experience as potentially explanatory variables, without proving that they *actually* account for the difference. The appeals court noted that in cases brought under Title VII of the Civil Rights Act, the plaintiff would have the burden of showing that the firm's explanation was a pretext; citing *Reeves v. Sanderson Plumbing Products, Inc.* 530 U.S. 133, 142-43(2000). Under the Equal Pay Act, the employee only needs to show a difference in pay for "equal work on jobs the performance of which requires equal skill, effort and responsibility, and which are performed under similar working conditions". An employer who asserts that the difference in pay is due to a "factor other than sex" has the burden of production and persuasion; citing *Corning Glass Works v. Brennan*, 417 U.S. 188,204 (1974).

Several other observations made by the appeals court are worth mentioning. Even if education and experience explain some of the difference in starting salaries, there is no reason they should explain *increases* in pay as they depend on job performance. Indeed, if men start at higher salaries due to more education but women and men perform similarly, women should receive more rapid pay raises and the salaries should tend to converge.

To illustrate the additional insight formal statistical analysis provides, we fit a regression equation to the final salary data. Unfortunately, the date when some employees left the firm was not reported, so their seniority is approximated by 2007 – year of hire +.5. The fitted equation, with standard errors below, is:

Salary = 70356 + 2396.17Seniority -31129Gender; R^2adj = .598
(1065.745) (7096.078)

Seniority is a significant factor, p-value - .038, however, the gender coefficient is highly significant, p-value = .0004. The estimated $31,129 differential women business managers received, after accounting for seniority, is clearly substantial for jobs with a median salary of about $70,000.

To ascertain whether there was evidence of salary convergence, we added an interaction term, measuring the joint effect of seniority and gender. A positive coefficient for the interaction term would reflect convergence, i.e. the pay of women was rising faster with time on the job. Fitting that equation estimated the interaction effect at -$1080.94 per year, indicating women's pay grew at a lower rate than males,

however, it was far from statistically significant (p-value =.667). Clearly, this result supports the 7th Circuit's skepticism of the explanation offered by the defendant as it shows no evidence that pay increases gradually reduced the differential between the pay of female and male managers.

It is also important to examine the role of gender in the setting of starting salaries. Since initial pay rates increase over time, we included the effect of seniority (when an individual was hired). The gender coefficient was -$34590, which was highly statistically significant (p-value < .0001).

When the case proceeds, more information on the education, prior experience and starting date of all employees presumably will become available for analysis as well as additional information on the date some of the managers in the data left the firm. This will allow a more careful analysis of the average yearly pay raises and a more precise estimate of seniority.

The data in Table 2 can be used to illustrate another important aspect of statistical criticism. Suppose the defendant were asked to justify the differential in 2007 or final pay and asserted that they had to offer men more initially. The defendant could submit the following regression fit to the data:

Salary = 8872.20 + .879 (Hire Salary) + 1637.59 (Seniority) -729.97 Gender

The equation is a good fit (R^2adj =.875.) and hire salary is highly significant (p-value < .0001), seniority is significant (p-value =0.16) but gender is far from significance (p-value =.909).

If one is concerned with the fairness of current or final salaries, initial salary is on the "causal pathway". Since the management of the firm under scrutiny determines both the initial and current salaries, it is statistically unsound to use hire salary as a predictor of current salary unless one can show it was determined fairly. This problem arose in the smoking and lung cancer controversy as some tobacco firms argued that one should compare the cancer rates of smokers and non-smokers who had other diseases related to smoking[19].

[19] See: GASTWIRTH, J.L. and GREENHOUSE, S.W. (1995), *Biostatistical concepts and methods in the legal setting*, in: *Statistics in Medicine*, 14, pp. 1641-1653 for other examples of this problem and the relationship between methods used to analyze some types of data arising in both biostatical and legal applications.

4. Discussion

Several of the cases discussed here show that courts have looked carefully at proposed explanations of statistically significant differentials in pay or promotion between minority and majority employees or significant disease rates between individuals exposed to a toxic agent and non-exposed. At other times, courts have simply accepted a proposed explanation of significant disparity without examining whether an analysis including information on the proposed explanatory factor does explain all or most of the differential or satisfy the conditions of an appropriate Cornfield type of inequality.

The *Dukes v. Wal-Mart*[20] case concerned the evidence required for plaintiffs to proceed as a class action. In particular they needed to show the firm had a policy that was *common* throughout its stores that negatively affected the pay and promotion possibilities of female employees throughout its stores. In the opinion, the majority of the court rejected plaintiffs' analysis of data obtained by pooling information about the employees of *all* stores in each of 41 regions, stating "A regional pay disparity, for example, may be attributable to only a small set of Wal-Mart stores, and cannot by itself establish the uniform, store-by-store disparity upon which the plaintiffs' theory of commonality depends.

There is another, more fundamental, respect in which respondents' statistical proof fails. Even if it established (as it does not) a pay or promotion pattern that differs from the nationwide figures or the regional figures in *all* of Wal-Mart's 3,400 stores that would still not demonstrate that commonality of issue exists. Some managers will claim that the availability of women, or qualified women, or interested women, in their stores' area does not mirror the national or regional statistics. And almost all of them will claim to have been applying some sex-neutral, performance-based criteria—whose nature and effects will differ from store to store".

Notice that because it is possible that a few stores could cause the pay and promotion disparities in a region, the Court did not require the defendant to produce evidence that this occurred. As noted in a reanalysis of the promotion data[21] the data is consistent with an overall system in which the odds of a female employee being promoted at a

[20] 131 S. Ct. 2541 (2011).

[21] GASTWIRTH, J.L., Bura, E. and MIAO, W. (2011). Some important statistical issues courts should consider in their assessment of statistical analyses submitted in class certification motions: implications for *Dukes v. Wal-mart*.*Law, Probability and Risk*, 10, 225-263 at pp. 237-245.

managerial level is only 80% of those of a male. Moreover, in 161 statistical tests made on the promotion data to four managerial positions in the 41 regions, which in a fair system would be expected to produce about 4 significant results disfavoring each gender[22], *none* resulted in a statistically significant shortfall of males but 110 resulted in a significant shortfall of females. It does not seem plausible makes that a few stores in each of the 41 regions could have created such an extreme pattern of disparities.

In fairness to the majority, the plaintiffs did not present either the analysis of the promotion data or analyses combining the results of defendants' store-wide regressions, which showed a highly significant general pattern of underpayment of women[23]. Combination methods allow the analyst to stratify the data into appropriate subgroups, e.g. working in the same unit of a company or having similar education and then assess whether there is a common pattern in most of the strata[24].

While the specific cases discussed in this paper came from the area of equal employment, the methodology originated in the analysis of

[22] Since courts use a two-sided .05 level test, the probability of finding a significant result showing a disadvantage to each of the genders is 025 or 1/40. When 161 tests are made, one expects one-fortieth or about 4 to show a significant result indicating a shortfall in promotions for each gender.

[23] See n. 21 at pp. 254-259.

[24] When there are many strata, e.g. the 41 regions in the *Dukes v. Wal-mart* case, in fair system, one expects a .05 level test to indicate a statistically significant difference in two regions. Similarly, if there is a "biased" system, e.g., the odds of promotion of females typically are only 80% of those of a male; the data in a few regions may be consistent with fairness. This is a result of the sampling variation, which in this context is not actually due to the data being a random sample (all the employees are included in the database) but due to the inherent random assignment of employees with different levels of ability and experience to the various stores. Thus, it is not reasonable to demand that plaintiffs demonstrate a common pattern in *every* store or more generally, strata before concluding that there is a common system that undervalues the qualifications of a protected group. Although In *Dukes v. Wal-mart*, the plaintiffs; expert included an indicator for each store when he analyzed the combined sample for each region, this is not the same as comparing the pay of employees in each store and then combining either the actual estimates of the store-wide disparities or the p-values of the regression analyses of pay in each store into an overall statistical test as described in Gastwirth et al, n. 23. The dissenting opinion in the decision apparently thought that these two analyses were equivalent.

epidemiologic studies and is applicable to data in many areas of scientific investigation. It should prove useful in the evaluation of "explanations" of increased risk of side effects from a new drug, of an increased incidence of a specific cancer from exposure to a toxic chemical in the workplace or residential neighborhood or of the poor performance of students in a school district[25]. It is important for judges and policy makers to be aware of these analytic tools, so that they can carefully evaluate studies submitted in support of a party in a legal case or an advocacy group in the context of a regulatory review. In particular, they should require critics of studies to demonstrate that an alleged flaw, e.g. an omitted variable or a small amount of missing data could plausibly affect the main conclusions of an otherwise sound study.

[25] A related method for combining the information in several epidemiologic studies is presented in GASTWIRTH, J.L. (2012). Should law and public policy adopt "*Practical Causality*" as the appropriate criteria for deciding product liability cases and public policy? (To be submitted to a special issue of *Law, Probability and Risk* from the Quantitative Justice and Fairness Conference in Lisbon, May, 2012).