Statistics in the Courtroom: Building on Rubinfeld

Richard O. Lempert

University of Michigan Law School, rlempert@umich.edu

Available at: https://repository.law.umich.edu/articles/2388

Follow this and additional works at: https://repository.law.umich.edu/articles

Part of the Evidence Commons

Recommended Citation


This Article is brought to you for free and open access by the Faculty Scholarship at University of Michigan Law School Scholarship Repository. It has been accepted for inclusion in Articles by an authorized administrator of University of Michigan Law School Scholarship Repository. For more information, please contact mlaw.repository@umich.edu.
STATISTICS IN THE COURTROOM: BUILDING ON RUBINFELD

Richard Lempert *

INTRODUCTION

As the use of statistics in litigation has burgeoned and as more complicated statistical techniques have entered the courtroom, concern for the way courts use statistics has mounted and efforts to instruct lawyers and judges on the wise use of statistics have begun. Professor Rubinfeld’s paper 1 is a contribution toward this end. Two ideas at the core of this paper are particularly important if we are to develop a more satisfactory approach to the use of statistics 2 in the courtroom. The first is Professor Rubinfeld’s caution against the talismanic use of the .05 level of significances as a test of what aspects of a statistical study are important to a legal factfinder. The second is his call for more attention to sensitivity testing than is customary in litigation research. 3


2. "Econometrics" is at once too fancy, too discipline bound, and too narrow to be the term of choice.

3. By litigation research I mean empirical research that has been specifically undertaken to address issues raised by the case at bar. Thus, if the X company is sued for sex discrimination, a study of the X company’s files to determine whether they systematically underpay their female employees would be litigation research. Expert testimony based on published studies about sex discrimination in the industry would report the results of social science research but not the subset of that research that I call litigation research.
of statistical significance to evaluate the importance of empirical research results to legal proceedings. In an earlier article, I noted that the values of social science are not the values of law, and that the .05 level reflects the social scientist's conservatism with respect to Type I error.4 This concern with minimizing Type I errors makes sense for social science purposes because independent replications, the best check against unreliable results, are seldom done. But what makes sense in social science may not in law. Statistical significance and substantive significance do not necessarily coincide; the likelihood of a statistically significant relationship varies both with sample size and the appropriateness of the statistical procedures. Moreover, the law may explicitly attach values to both Type I and Type II error that have important implications for the level of significance we should demand.

Professor Rubinfeld largely shares these views. He acknowledges that both more and less conservative standards than the .05 level can be appropriate on occasion. However, like most good statistical experts who discuss the probative value of statistical evidence, he appears more concerned with the mistake of giving weight to statistical evidence that has little probative value than he is with the mistake of dismissing statistical evidence that does not achieve the .05 level yet is reliable.5 The law, however, seems to have made just the opposite determination. The Federal Rules of Evidence define as relevant all "evidence having any tendency to make the existence of any fact that is of consequence to the determination of the action more probable or less probable than it would be without the evidence."6 Surely statistical evidence that is significant at the .10 level or even the .50 level often meets this test.

The Federal Rules of Evidence nonetheless suggest two grounds for excluding evidence that does not reach some minimal level of statis-

   
   In the history of the development of statistical methods .05 was most emphatically not regarded as "conservative." To R. A. Fisher, anything not significant at the two-sided 5% level is not entitled to credence as evidence at all. "Significant at the 5% level" was for him merely suggestive, "significant at the 2% level" was plausible, and only "significant at the 1% level" could be considered convincing. I share his view. Remember that real effects don't usually hover at the border of statistical significance. Most of them are statistically unambiguous (as is most evidence of discrimination).

5. Interestingly, the former mistake is analogous to Type I error and the latter to Type II—perhaps what we see in these analyses is simply the statistician's habit of paying more attention to Type I than to Type II questions carried over to policy analysis. My characterization of Professor Rubinfeld's views applies more to the draft version of his article, see Rubinfeld, Econometrics in the Courtroom (Oct. 1984) (on file at the offices of the Columbia Law Review) [hereinafter cited as Rubinfeld Draft], than it does to the current version.

tical significance. First, Rule 703 provides that an expert may rely on evidence not otherwise admissible, but only if the evidence is "of a type reasonably relied upon by experts in the particular field in forming opinions or inferences upon the subject . . . ." To this we may add the test of Frye v. United States, still treated by some courts as good law, which holds that scientific evidence will not be admitted unless its scientific validity is generally acknowledged by professionals working in the appropriate scientific field. Arguably, experts of various sorts (be they from statistics, sociology, economics, or any other field) would neither rely on statistical results that fail to reach the .05 level of significance nor acknowledge the validity of conclusions based on statistical evidence that does not reach this level, so under either Rule 703 or the Frye test the evidence could be excluded as an unfit basis for expert testimony. These evidentiary rules should not, however, be interpreted in this way, for as I have already pointed out the values leading social scientists to establish .05 as the conventional mark of statistical significance are not those of law. Not only do social scientists have reasons to be especially concerned with Type I (as opposed to Type II) error, but scientists have the luxury of withholding judgment on an issue, a luxury which courts lack. For example, a careful scientist who fails to reject a null hypothesis because the coefficient on the crucial variable differs from zero at only the .15 level would not thereby conclude that the null was true; rather, he would reserve judgment on the issue. Courts, by contrast, typically must decide between two competing alternatives. If the law fails to reject its equivalent of the null hypothesis because the evidence pointing in this direction is not sufficiently convincing, it effectively regards the null as established. Thus, while statistical scientists properly determine the acceptable ways of conducting statistical tests, courts should not defer to conventional scientific judgments about when statistical evidence merits attention.

The second possible ground for barring statistical evidence that has not met some significance level criterion is Rule 403 of the Federal Rules of Evidence, which provides that relevant evidence may be excluded if its probative value is substantially outweighed by its tendency to waste time, confuse the issues, or mislead the jury. Clearly the ten-

8. 293 F. 1013 (D.C. Cir. 1923).
10. Frye, 293 F. at 1014.
11. This does not mean that statistical experts have nothing to offer on the issue of what to make of evidence that fails to achieve (or indeed exceeds) a conventional level of statistical significance. Their expert views on the implications of such evidence may provide helpful or even essential guidance. But if this is to be the case, the expert's discussion must reflect the context in which the evidence is offered and the values involved rather than decontextualized conventional wisdom. The root of the problem is, I think, that courts are engaged in a decisionmaking task for which a Bayesian approach to statistics is more appropriate than a frequentist one. See infra text accompanying note 15.
dency of statistical evidence to prove the point on which it is offered\textsuperscript{12} diminishes with declining statistical significance, and the likelihood that it will waste time increases.\textsuperscript{13} In addition, statistical evidence based, as it is, on unfamiliar mathematics has a natural tendency to confuse nonmathematically trained factfinders. This confusion arguably increases as the statistical significance of the evidence declines, since the possibility that the evidence will be overweighted if any attention is given it increases.

One problem with the Rule 403 objection is that any statistical evidence carries with it a risk that it will be overweighted or otherwise confusing to the jury.\textsuperscript{14} Indeed, calling evidence that reaches the .05 level "significant," even if the qualifier "statistically" is added, may lead to confusion or overweighting, since lay factfinders are likely to find it difficult to separate statistical and substantive significance, although there is no simple association between the two.

More fundamentally relevant evidence should be excluded only as a last resort if confusion is the problem. The first resort is to attempt to clear up the confusion. One possibility is that information pertaining to significance levels might be excluded. Arguably, for litigation purposes (and most other decisionmaking tasks), an explicitly Bayesian approach should be taken.\textsuperscript{15} To pursue the issues that this possibility raises would, however, involve us in a debate whose merits I am not going to rehearse here. Let me simply note that the Bayesian perspective is unlikely to transform the way statistical data is evaluated for litigation purposes at any time in the foreseeable future. Economists, sociologists, and others will continue to run their significance tests, and they will continue to inform factfinders of the results.

A second, less radical alternative, which I espouse, is to change the

\textsuperscript{12} I am assuming in this discussion that one party offers statistical evidence to prove a fact and the other party argues that the evidence should not be received or does not tend to prove that fact because while the evidence points in the direction of the fact it does not achieve statistical significance. It is also possible that a party contesting a fact will seize on the insignificance of a particular result to argue that a contested fact does not exist.

\textsuperscript{13} Indeed, it may be that at a certain point the relevance of statistical evidence is unknown. For example, in an equation testing for sex discrimination, is evidence that the coefficient on the sex variable is in the predicted direction, but significant at only the .30 level, more consistent with discriminatory or nondiscriminatory defendant behavior? If the relevance of evidence is so unclear that the jury can only engage in ungrounded speculation about what it implies, that evidence may be excluded under Fed. R. Evid. 402 as irrelevant even though, properly understood, the evidence would have some tendency to prove or disprove a fact in issue. Lempert, Modeling Relevance, 75 Mich. L. Rev. 1021, 1029 (1977).

\textsuperscript{14} I am not concerned with underweighting since this would give the opponent of the evidence no cause to object.

\textsuperscript{15} For applications of Bayes Theorem, see, e.g., Finkelstein & Fairley, A Bayesian Approach to Identification Evidence, 83 Harv. L. Rev. 489 (1970); Lempert, supra note 13, at 1022–32.
way that juries are informed of the results of significance tests whenever
data appear to have evidentiary value. If the significance or nonsignifi-
cance of a variable is truly important to the litigation, the factfinder
should not be told of a significance level and then left either to be awed
by the variable’s importance (if it is significant) or to dismiss it (if it is
insignificant) or to try to puzzle out what the significance level really
means. Instead, the implications of a significance level for Type I and
Type II errors should first be explained to the factfinder. Then, the
statistical evidence should be presented both in terms of what the
strength of the relationship implies for the possibility of these two types
of error and in terms of the plausible substantive implications of the
variable, given the level of statistical significance.16

The analysis to this point is consistent with Professor Rubinfeld’s
suggestions. I part company from him, however, when he suggests that
levels of significance be chosen with an eye toward litigant or expert
behavior.17 As an example of where this perspective can lead and the
problems with it, consider Rubinfeld’s suggestions in an earlier version
of his paper that we choose a low significance level “when the cost of
litigating certain types of cases is likely to be high for all prospective
litigants,” and to compensate for the possibility that experts will en-
ga in model searching.18 I find both the value and empirical judg-
ments that are implicit in these suggestions questionable.

Litigation that requires statistical expertise is already expensive.
One would expect this fact alone to be a substantial disincentive to liti-
gate. I fail to see why a further disincentive should be established for
such cases. Indeed, fee-shifting statutes have been passed because we
think that certain kinds of expensive litigation raise issues that should
be brought to court.19 If we think that other kinds of expensive litiga-
tion are not worth the cost, let us change the substantive law to pre-
clude such battles entirely. Or, if that is more discouraging than we
wish to be, let us raise the burden of proof directly for all cases rather
than raise it indirectly and haphazardly for cases that rely on statistical
evidence.

In addition, only plaintiffs are likely to be discouraged from litiga-

16. This should probably be presented in terms of a range of values. For example,
after explaining the 95% confidence interval, the judge might say, “If the model used is
correct, we can conclude with this degree of confidence that being a woman costs a clerk
at company X between $.16 and $.96 an hour in wages.”

17. See Rubinfeld, supra note 1, at 1063. The argument was considerably stronger
in the draft version of Professor Rubinfeld’s paper which stimulated these reflections.

18. See Rubinfeld Draft, supra note 5, at 18–19. There are echoes of Professor
Rubinfeld’s views in the current version, see Rubinfeld, supra note 1, at 1063, but
enough has changed that my discussion should not be seen as criticism of Professor
Rubinfeld. Instead I am using his earlier, tentative suggestions heuristically, since his
earlier views raise interesting issues that merit discussion.

fees to the prevailing party in a Title VII suit).
tion by the suggested change. With statistical evidence barred because it meets, say, only the .05 rather than the .02 level, defendants might fight cases that they would promptly settle if the statistical evidence were admissible. The more general point is that in thinking about the incentive effects of rules regarding statistical evidence on litigant behavior, one cannot assume that the statistical evidence is either the only or the most important evidence in the case. Thus the effects of changing the threshold of admissibility for statistical evidence are not obvious.

To change the level of required significance to prevent model searching would, of course, penalize those experts (or the clients of those experts) who do not search. Indeed, it might encourage model searching. The expert who specifies in advance a well thought-out model and finds that the crucial coefficient is significant at the .05 level when the court, to prevent model searching, has said it would not treat a finding as important unless the .02 level is reached will, no doubt, proceed to examine less well thought-out models in the hope that in one of them the crucial coefficient will reach the .02 level. Indeed, the expert might not feel this is unprofessional since the court has set the more stringent level to take into account model searching behavior.20

If there is a problem with model searching, and I agree with Professor Rubinfeld that there is, there is a better solution. It is the one he identifies.21 This is to require the expert to disclose all model searching conducted. It also requires that lawyers learn to question experts on discovery about model searching and that lawyers learn what this implies for attained significance levels and how to communicate this information to triers of fact.22

II. SENSITIVITY TESTING

With respect to sensitivity tests, I generally endorse the approach

---

20. In the general case the more stringent level may mean that evidence from the Nth model tested, which is the first to achieve statistical significance, is more probative than the evidence from Model "N-M" which is the first to attain some less stringent level of statistical significance. But this is not necessarily true. First, it depends on the magnitude of "N-M"; that is, on how many more models are searched in an effort to achieve significance under the more stringent standard. Second, the evidentiary value of a coefficient significant in the predicted direction at the .05 level in a well thought-out model may be greater than the evidentiary value of a coefficient significant at the .01 level in a poorly considered model.

21. See Rubinfeld, supra note 1, at 1073, 1095.

22. While I am giving Professor Rubinfeld credit for a solution, I should also give him credit for what may be the strongest argument against manipulating significance levels to affect litigant or expert behavior. This is that there is no constant relationship between statistical and substantive significance. Thus, to manipulate significance levels to achieve the desired end would be to eliminate more substantively important but less statistically significant evidence in one case, while allowing in more statistically significant but less substantively important evidence in another.
taken in Professor Rubinfeld's article. I have only a few brief ideas to add. The first is that sensitivity testing can be for those fighting the implications of statistical evidence something like model searching is for those who offer it. One seeking to dismiss the results of a statistical analysis can explore different models until he or she discovers a variable or a transformation or a technique that reduces the crucial coefficient in the opponent's analysis to statistical insignificance and then can tell the court that the opponent's analysis is useless because it is sensitive to an omitted variable, functional form, the presence of outliers, or the like. But demonstrating such sensitivity does not show that the opponent's work is useless. If the opponent's model is theoretically sound and the alternative is theoretically implausible, the fact that the model is sensitive to the selected specification is no reason to suspect its validity. Even when the alternative specification is theoretically plausible, if the original specification is well thought-out, the sensitivity analysis does not totally undercut the opponent's statistical case. It merely suggests the need for caution and, more valuably, directs the factfinder's attention to a crucial theoretical disagreement that must be resolved to evaluate the competing claims.

The legal community could use explicit guidance from econometricians and others about when sensitivity to basic specification decisions is likely to seriously undercut the implications of particular regression results and when it is likely to be less important. In giving such advice, however, statistical experts must hold in mind that the law's values may differ from scientific ones. Consider, for example, an equal payment sex discrimination case in which the plaintiff's regression analysis is sensitive to several outliers: women who, despite relatively high education and experience, earn rather low incomes. Omitting these outliers might reduce the negative impact of sex on earnings to a point where it is no longer statistically significant.23 A social scientist aware that some unmeasured variable may explain the outliers' positions might argue that omitting them provides a "truer" picture of the likelihood that the company discriminates on the basis of sex. Yet, it may be that the outliers are outliers precisely because they were victims of sexual discrimination.

To eliminate the outliers because it is good social science to be wary of outlier-dependent regressions may be to eliminate what is properly among the plaintiff's strongest evidence in the case.24 For example, consider the process of setting a Bayesian prior to bring to the analysis of data suggesting that, other things being equal, women in a high-tech company are paid on the average a little bit less than men of the same education and experience. One might set the prior quite low

23. See Rubinfeld, supra note 1, at 1070.
24. It is the wariness and not the solution that is good social science. This wariness should lead to an attempt to understand why outliers exist and thus to deal with the data appropriately. Arbitrarily ignoring outliers is no solution to the problems they pose.
on the theory that modern companies in competitive environments have no history of discrimination and little to gain from it. Now suppose that it can be shown from independent evidence that at least three women in the company had been dramatically discriminated against because of their sex. Clearly, one would raise one’s prior after learning that the company discriminated in some cases, and it becomes more reasonable to suppose that a male bias is pervasive.

Of course, in the typical regression analysis, one would not know why the three women are outliers. To eliminate them is to assume either that their position is not the result of sex discrimination or that, if it is, their situations are so unique that their treatment has no bearing on how to interpret the data pertaining to the pay of other women. The alternative, of course, is not to avoid a sensitivity analysis. Rather, it is to argue that the implications of the analysis should be discussed carefully in light of the legal values at stake. In particular, it is to emphasize that “cookbook” approaches of all sorts, even if accepted as general rules for social science, must be rethought in the context of particular cases and the legal and value issues they involve.

My second point is closely related. Sensitivity testing is a way of evaluating the risks in relying on the implications of a particular statistical analysis. If crucial regression results are sensitive to functional form or the presence of an omitted variable, we rely on the results at our peril, for if the reality we seek to portray is more properly captured by an alternative specification than by the original model, the original results may offer a substantially distorted picture of what is occurring.

This suggests that we search for models which on theoretical grounds are sufficiently compelling that we are justified in relying on them even if related specifications yield rather different outcomes. As social scientists trying to advance social theory this is difficult. Rarely, if ever, do we know enough about the world we are investigating to say that it is no cause for concern when a plausible alternative specification yields results that are substantively different from those of our original model. For example, if we are trying to understand how salaries are determined in company X, we might test a regression model that includes variables for sex, experience, supervisory responsibility, hours worked, job classification, and performance ratings, but not education. If the coefficient on the sex variable were significant and suggested that women were meaningfully disadvantaged, we might provisionally conclude that company X discriminated against women. However, if adding education as a variable reduced the sex coefficient to the point where it was no longer significant, we would probably conclude that the worker’s sex was not important to an understanding of how salaries

25. A daring defendant might argue that such evidence tends to show that the other women employees were not discriminated against, because when this company discriminates the effects are clear to all.
were set, although education did matter. Alternatively, we might argue that the company did discriminate on the basis of sex, but that the relationship between sex and education in the company’s workforce was close enough that including education as a variable suppressed the effect associated with sex. On statistical grounds, we could not choose between these explanations.

In litigation, however, the law’s normative system may allow us to specify models that are for the purposes of the case at hand not vulnerable to this kind of indeterminacy. For example, suppose we ask the management of company X what variables they consider in setting salaries and they list all the variables we mention above except sex and education. We ask, “What about education?” and are told “It’s not considered. We’re interested in what workers do on the job and not in how much book learning they have.” Then we run our regressions and find that sex apparently matters when added to the other variables we have identified, but not if education is added as well. The sensitivity of our model to education arguably has no implications for the reliability of the results suggesting sex discrimination. The company has told us explicitly that education is not considered, and we have no good reason to believe that it is.26

Alternatively, education may be a proxy for such factors as performance ratings, which the company does not deny taking into account. But if we have adequately measured all these factors the value of the proxy is lost. Even if we have not measured all such factors, there remains the question of what the company can measure. If, for example, education affects job performance in ways not captured by our variable “performance rating,” this will not threaten our conclusion of sex discrimination if bureaucrats set salaries within the company, and their only available measure of performance is performance rating scores.27 Thus, in choosing between model specifications for the purposes of liti-

---

26. The company has also told us that sex is not considered, but we have reason to believe it is. It is also possible for the company to discriminate without knowing it, and, knowing or not, the company has a substantial incentive to deny any influence of sex. Furthermore, it may be appropriate to hold the company to what it says it is considering since we expect salaries to be set according to a rational policy of which the company is aware. We are doing this when we hold it accountable for the effects of sex discrimination, which is illegal, and do not consider the possibility that salary setting on the basis of education, which is legal, accounts for the lower salaries of female employees.

27. The law’s norms may even, on occasion, justify a specification that we think is mistaken. Suppose we have a company that employs large numbers of black women relative to the number of blacks and women in its workforce. A model without race included may suggest sex discrimination, but once race is included the effects of race may be strong and those associated with sex may disappear. As social scientists we may be reasonably confident that we have a case of racial and not sexual discrimination. Yet in litigation a company may not be allowed to or may not want to defend a sex discrimination claim on the ground that it is really engaging in racial discrimination, even if the result of disallowing that defense will be the award of damages to some white women who were not in fact discriminated against.
gation, there may be grounds for preferring one to another that would not be available or would be more tenuous if one were simply seeking a scientific understanding of what is going on. One implication is that the law may misunderstand the real situation. The law, however, is in the business of holding people responsible, and it is not only not obviously wrong as a matter of morality but it is clearly permissible as a matter of law to hold people responsible for what they think and say they are doing even if some otherwise plausible statistical tests suggest that the reality is more benign than the party's own standards suggest.

It follows from this position that experts should work closely with lawyers in designing statistical models, for it may be possible to design models that are legally more robust than sensitivity analyses might indicate.

Finally, there is a point that applies both to model searching and sensitivity analyses. Professor Rubinfeld properly calls for experts to reveal both the models they have tested and the sensitivity testing they have done. But such calls are not self-enforcing. In addition, it is not clear what should count as model searching or sensitivity testing. Experts commonly look at their data from various perspectives in deciding how best to approach the analytic task. What exactly should be revealed?

Ideally it might seem that experts should be required to keep all hard copy they generate in the course of their various analyses and make it available to the opposition. But this "bright line" test may lead to two types of problems. First, it might simply induce experts to print out less hard copy. Results would be inspected on the screen, but unless they aided the client's case, hard copy would not be generated. On the other hand, if every test analysis was preserved and transmitted to the opposition, the opposing side might make far more out of casually done exploratory testing than the data merit. In the hands of a deft and unscrupulous cross-examiner the results of quick and casual exploratory testing might easily be blown up in importance, and the factfinder might be led to believe that the expert's theoretically sound, well thought-out tests were contradicted by other work the expert had done and were designed to hide earlier adverse results.

I do not have any easy answer to these problems. I expect that how well we do in dealing with them will turn in large measure on the ethics of expert witnesses, the ethics of attorneys who employ and cross-examine them, and the ability of attorneys both to understand what statistical experts are doing and to communicate that understanding to triers of fact. I also expect that if we want experts to act ethically and disclose the model searching and sensitivity testing they have done, it would help if we had a clear statement of exactly what activities tending in these directions should be disclosed. It might also help if we had some system for periodically reviewing the behavior of experts to determine
whether the standards set had been met and perhaps to sanction experts and attorneys for willful failures to comply.

Thus far, I have attempted to build on themes that Professor Rubinfeld develops in his paper. Now I would like to suggest two themes that are largely missing from his paper which, in my opinion, figure importantly in his mission of helping us think more clearly about the use of statistics in litigation. They are, first, that when statistics are used in court, they are being employed in the context of an adversary system, and second, that statistical experts, like all important witnesses, must attend to how their testimony fits into the story that the client is attempting to tell. A consideration of these themes qualifies portions of the message that Professor Rubinfeld conveys.

III. An Adversary System

Litigation, as experts are often all too painfully aware, is in this country an adversary proceeding. This fact has implications for almost every aspect of expert testimony and bears importantly on such matters as the ethical standards that should be demanded of experts, the best modes of presenting statistical testimony, and the desirability of employing “neutral” court-appointed experts. In this Article, however, I shall attend to only one way—indeed, to only part of one way—in which the existence of an adversary system might affect our analysis of what to expect of statistical experts in court.

In an adversary system, failure to respond to an opponent’s argument is itself an important piece of information. This means that the statistical standards that we might impose on a social scientist advancing a scientific hypothesis may be less necessary or even inappropriate in the context of litigation. This is particularly likely when the information costs of the two parties are markedly different.

To illustrate this by example, consider Professor Rubinfeld’s discussion of Melani v. Board of Higher Education, and his suggestion that on a department by department basis, male and female faculty at City University of New York (CUNY) might have been paid the same, but the average salary of female faculty might have been substantially less than that of male faculty because women may have been clustered in departments (like education) which, because of external market forces, have low salary scales, while men were clustered in departments (like engineering) that have higher ones. Professor Rubinfeld suggests that the plaintiffs in Melani may be criticized for failing to control for departmental appointment in their regression analysis. Professor Rubinfeld’s critique would clearly apply if the study in question was an ordinary social science investigation. Surely, we would want a social

29. See Rubinfeld, supra note 1, at 1069.
scientist seeking to determine why CUNY's female faculty were earning less on the average than their male counterparts to consider the possibility that women were clustered in departments with historically low salary scales. But in the adversary context, it may be less important to demand this. If the defendant fails to pose the possibility that departmental status might explain apparent sex-based discrimination or, if having raised the possibility, the defendant fails to test for it, the defendant's failure is itself reason to believe that a consideration of this plausible nondiscriminatory explanation would not exonerate the defendant. The case for resting the burden of introducing plausible alternatives on the defendant is particularly strong if, as will often be the case, the defendant is in a better position than the plaintiff to identify plausible nondiscriminatory explanations for apparent discrimination. Arguably CUNY was in this situation, for at least at the outset of the litigation, they must have known better than the plaintiffs how women faculty were distributed across the University and how market forces affected the salary structures of different departments.

The problem is general. It also applies to Professor Rubinfeld's suggestion that it is often desirable to test the probability of evidence under the null hypothesis of no discrimination against the probability of that evidence under some specifically hypothesized mode of discrimination rather than against the vague alternative of discrimination in any form. As Professor Rubinfeld's example demonstrates, his preferred approach can aid either the plaintiff or the defendant depending on the data and the alternative posed. Ordinarily this is reason enough to put the burden of advancing specific alternative hypotheses and testing their plausibility vis-a-vis the null hypothesis on the party likely to benefit from this approach. If this party chooses not to proceed in this fashion, it is fair to assume that the procedure either is not appropriate in the given case or that it would not substantially alter the import of the other evidence.

One can, however, push this argument about the adversary system too far. Complexities arise, especially when multiple regression is an appropriate methodology, and the question is which party is responsible for evaluating the effects of which variables. To draw on an example that Professor Rubinfeld used in an earlier version of his paper, assume that the proportion of women hired for a particular position is substantially less than the proportion of women in the company's applicant pool and that the job in question requires substantial education as well as previous job experience. In these circumstances the disproportionate underrepresentation of women among the company's hires is, at best, only weak evidence of discrimination. It may be explained by the fact—if it is a fact—that female applicants are less experienced or

30. See the discussion in Rubinfeld, supra note 1, at 1055-59.
31. See Rubinfeld Draft, supra note 5, at 16.
less well educated than their male counterparts. Who should bear the burden of testing this possibility? Many would unthinkingly place this burden on the plaintiff. I would argue that given the law and the adversary system, this placement is usually but not always correct.

First, consider the situation where the defendant has immediate access to the data needed to test nondiscriminatory alternative hypotheses (i.e., that experience or education explains the underrepresentation of women), and will not provide the plaintiff with all relevant information, or will only provide relevant information in a way that promises to make the analysis unnecessarily expensive. In these circumstances, I think the burden of showing the plausibility or implausibility of the employee value hypothesis properly rests on the defendant, and if the defendant presents no data disproving the plaintiff’s contention, the factfinder may properly conclude that the data would not have disproved it.

The situation is arguably the same where the crucial data is equally accessible to the plaintiff and the defendant or where the defendant will cooperate in giving the plaintiff access to the data. What is to be lost by forcing the party claiming nondiscrimination, rather than the party claiming discrimination, to introduce evidence on the plausibility of alternative hypotheses? Arguably, putting this burden on the defendant will give a greater “nuisance value” to the plaintiff’s claim, and the defendant may be willing to settle for more than the plaintiff deserves to avoid the expense of the analysis that may defeat the plaintiff’s case. But, by the same token, the plaintiff who must test the plausibility of the alternative hypotheses may settle for less than she deserves in order to avoid the expense of the alternative hypotheses analysis that might strengthen her case. I see no basis for choosing between the parties on this ground. Moreover, if, as will usually be the case in employment discrimination litigation, the defendant has better access to the data needed to test the “job qualification” hypothesis and knows more about how the data have been coded and computerized, forcing the plaintiff to test for the rival hypotheses will introduce unnecessary transaction costs.

Nevertheless, where the data are equally accessible, I think the plaintiff should be responsible for first testing the effects of certain variables, like measures of education and experience, that may explain an apparently discriminatory hiring pattern. The reason is that the law’s normative model of discrimination does not label as discriminatory those situations in which women are less frequently hired or paid less than men. The label is limited to situations in which women are disad-

32. The defendant might have the data on employee education and experience on tape for internal management uses, but rather than provide the tape, he delivers a truckload of personnel files for the plaintiff to code.

33. By this, I mean the plaintiff’s expected damages times the probability the plaintiff will prevail.
vantaged for reasons that have nothing to do with permissible predictors or measures of a person's likely or actual job performance. Simply demonstrating that women are paid less than men does not show this. Nondiscriminatory explanations for gross hiring or wage disparities between the sexes are in the general case sufficiently plausible that the plaintiff should ordinarily be required to evaluate a model that contains a reasonable set of plausible predictors of job performance in order to establish a prima facie case on the basis of statistical evidence. This obligation does not, however, extend as far as the scientist's obligation ideally extends. So long as the plaintiff's model contains a reasonable set of plausible predictors, the burden should be considered met. If the defendant believes that other rival hypotheses are important, the defendant properly bears the burden of collecting the data and running the tests necessary to support other exonerative theories.

Ultimately the scope of the plaintiff's obligation to specify and test hypothesized alternatives to the null hypothesis that do not entail discrimination (e.g., that wage disparities reflect job-related educational, rather than sexual, status) is and should be set by legal, not statistical, norms. Nevertheless, since both law and statistics are usually sensible disciplines, the standards of both will often point in the same direction.

IV. TELLING A STORY

Professors Bennett and Feldman argue that a trial is, in essence, a

34. The word "permissible" is important, and serves to emphasize the difference between models designed to test legal theories and those designed to test social science theories. Assume, for example, that sex is an important predictor of stockbrokers' sales because people are generally more willing to buy stock from men than from women. Nevertheless, a correlation between sex and sales would not justify setting women brokers' initial salaries lower than those of men. To do so in anticipation of future sales would be considered sex discrimination.

35. The obligation might be excused or pared down if the necessary data are unavailable or unduly expensive to analyze, at least if those conditions are due to the defendant's noncooperation.

36. To the extent that "legitimate predictors" explain apparently invidious discrepancies, this placement of the burden is likely to lead to a failure to bring suits or an earlier end to litigation that, from the plaintiff's point of view, would in any event have floundered after the defendant's statistical analysis. An early end to litigation that was not going anywhere should save both plaintiffs and defendants money, although in some cases particular plaintiffs may be worse off, for they may be confronting defendants who would settle on the basis of gross evidence of discrimination without doing their own controlled statistical analyses.

37. I assume the plaintiff has not generated this set by discarding variables which when included tended to support the rival hypothesis.

38. I would put in this category the theory in Melani that departmental affiliation explains sex-linked wage disparities, but I would not fault a judge who decided the other way. One might reasonably argue that this explanation is so plausible, easy to test, and obvious that the plaintiff should have borne the burden of considering the effects of departmental status.
 Individual witnesses tell their own stories, which the proponent of their evidence tries to weave into a larger, coherent whole and the opponent tries to negate. Ultimately, the party that prevails will be the party whose story makes the most sense to the factfinder. Statistical experts, like other experts, should seek to tell their stories in ways that make sense to the trier of fact and fit in with the stories of other witnesses to form a convincing whole. In particular, an expert should remember that however convincing he or she finds the statistical evidence after months or even years of being immersed in the data, the statistical evidence is unlikely to be the whole story. Where a factfinder rejects what is to the expert a convincing statistical case, the constructive reaction is not to condemn the factfinder's blindness, prejudice, or ignorance. Rather, it is to ask how the story could have been better told. Often the exercise will reveal that the factfinder acted rationally, given the way the story was told. Thus, Professor Rubinfeld gives sound advice, both technically and as a matter of persuasion, when he cautions that "[e]valuation of statistical results cannot be accomplished by the use of any single test statistic. Ideally, findings should allow the trier of fact to make independent inferences based on information relating to both statistical and practical significance."40 His example of translating a statistically significant coefficient on the sex variable in a wage discrimination case into an expectation of lost wages is also sound, for it is a way of helping the factfinder make sense of an otherwise difficult to understand statistic that fits nicely into the story that one party wants to tell.

This, however, is an easy example. It is statistically and literally an exercise in translation, and as such it may not alert us to more subtle problems that can arise. Consider, for example, Professor Rubinfeld's detailed discussion of the work he and Peter Steiner did in the ampicillin case.41 It is an impressive piece of work,42 yet, however appropriate and fine the statistical analysis is on its own terms, as a story something is lacking. To better explain what I mean, let me pose as a typical factfinder. In this guise, I do not know very much about the drug business, but I have noticed in shopping that generic brands of drugs are generally priced lower, and sometimes much lower, than their name brand counterparts. I also do not know much about economics, but if there is one lesson I recall from my introductory economics course, it is that market competition lowers the price of goods. Now Professors

40. See Rubinfeld, supra note 1, at 1068.
41. See Rubinfeld, supra note 1, at 1078–87.
42. My major reservation is that I think the model should have included detailed information about when drugs that compete with ampicillin entered the market and how their prices changed over time. While the time variable may be intended as a proxy for such effects, I am not satisfied, both because of the likely looseness of the fit and the fact that the price information on competitive drugs should have been available.
Rubinfeld and Steiner tell me that excluding generic drug companies from the ampicillin market did not noticeably increase the prices that governmental agencies paid for the drug. The net result is that I, as a typical juror, am impressed by Professor Rubinfeld’s careful analysis and awed by the credentials that he and Peter Steiner present, but totally apart from the plaintiff’s competing analysis, I am not convinced.

Nor is my reaction completely irrational. There is a good statistical defense that may be made. In Bayesian terms, my familiarity with generic drugs and the dynamics of competition means that once I learn that numbers of generic drug companies were excluded from the ampicillin market I rationally set a high prior probability on the likelihood of some price effect, so high that even after I learn of and understand the defendant’s statistical results, I nonetheless rationally conclude that the plaintiff has proved its case by a preponderance of the evidence.

The defendant’s story in the ampicillin case was “There was no effect.” That’s not much of a plot. What the defendant should have done to make sense of the statistical results was to explain why, in the context of the wholesale drug market of the late 1960’s, additional competition from generic houses should not have been expected to affect substantially the price of ampicillin sold to cities, counties, and states. With this information I would have revised substantially my Bayesian prior, the statistical analysis would have made more sense, and I would have decided for the defendant. To put the point another way, econometricians should remember that one of their comparative advantages in litigation (vis-a-vis statisticians, for example) is that they are also economists. Econometricians analyze data. Economists tell stories.

V. Lex Regis

I have touched on a number of issues in this commentary, although I have by no means exhausted the potential for building on Professor Rubinfeld’s paper. If there is any general theme to my comments it is Lex Regis: when econometrics or any other statistical specialty enters the courtroom, the law is king. The law’s norms and values ultimately determine what models are appropriate, what questions should be asked of data, how burdens should be allocated, and what various statistical results imply.43 Let me return one last time to the earlier draft of Professor Rubinfeld’s paper to illustrate this point. He gives an ex-

43. I am not saying that statistics cannot or should not inform the law of what good statistical analysis entails or how a statistical expert would interpret particular results. A strength of the law which we see in its increased receptivity to and understanding of regression analysis and other modern statistical techniques is that like a good ruler it is educable. I am only saying that when legal standards clash with or reinterpret statistical standards, it is the law which does and should prevail.
ample of twenty firms in an industry, the hiring practices of which are being scrutinized by employees who feel they were wrongly denied offers of employment. He stipulates that the firms hire independently of each other and do not consider sex or race. He suggests that it would not be surprising to find that a careful statistical analysis of the hiring practices of the firms leads to the conclusion that one of the firms has been guilty of discrimination. The context is novel, but the point is familiar and sound. If one looks at twenty independent random samples for an effect, it is not unusual for one to find a sample in which the effect appears at the .05 level simply as a result of random variation. Yet extending this proper statistical caution to the legal context is nonetheless problematic.

I have two problems with the example involving statistical issues that I shall only note in passing, and a third that is more intriguing and makes the point I want to emphasize here. First, I simply note that in the real world where we cannot know by assumption that companies are not discriminating, there is likely to be much less danger that a careful statistical analysis will, because of chance factors, mistakenly suggest discrimination. Even if companies hire independently, investigations into discrimination are not started, and are certainly not pursued to the point of a trial, at random. It is an empirical question, but I expect that before a case reaches the point of a substantial investment in a careful statistical analysis (often at the expense of the plaintiff's attorney), it appears to someone that there is a high prior probability that the defendant company has been discriminating. Thus, the danger that we shall label random processes discriminatory is likely to be substantially less than the analogy to the familiar paradigm of twenty independent tests suggests. But Professor Rubinfeld avoids this problem by stipulating that all twenty of his firms are innocent, and so long as we bear in mind the relationship of the example to the real world, I have no qualms about accepting the stipulation.

My second difficulty with the illustration is that I still wonder where the random element comes from. If it exists because sample surveys were done at each company, I accept Professor Rubinfeld's point fully, but the problem is no longer interesting in that there is an easy cure which is to look at the population data. If population statistics were examined in each company, some would question whether significance tests are appropriate at all. However, I will assume with Professor Rubinfeld that significance tests are useful as an indicator of whether relationships suggesting discrimination might plausibly result from the essentially random effects of unmeasured variables, and that it is appro-

44. See Rubinfeld Draft, supra note 5, at 20–22. The current version has an analogous example, see Rubinfeld, supra note 1, at 1064, but it does not illustrate my point as well. The statistical criticisms I make of the example I draw from the earlier draft also do not apply to the current version because the argument is not pushed as far.
priate to treat this situation as akin to the randomness associated with probability sampling.

This brings me to my third point. Assume that the statistical analysis Professor Rubinfeld posits has been carefully done, in that all obvious job-related variables have been well measured and included in the regression equation. Consider the situation Professor Rubinfeld has described. Women are substantially underrepresented in a company's workforce. The underrepresentation cannot be explained by the composition of the applicant pool or by the job-related credentials and skills of the women applicants, for we included measures of these in our model. Instead, if we could only know, we would find that—purely by chance—women (as compared with men) were more often interviewed by those employment officers least willing to hire; were more likely to apply for work when business was slack, more often had their interview disrupted by phone calls to the interviewers, and the like. In short, the business was blameless. Yet it does not follow that the law should listen to the "randomness" defense that exonerates it.

Assume, for example, that the applicable statute forbids discrimination, that discrimination ordinarily implies intent, and that the defendant can show that in some recent period twenty investigations into the hiring practices of area firms were started and only the investigation into the defendant's hiring practices yielded evidence of discrimination. I concede that on these facts a social scientist should not conclude that the firm was discriminatory in its hiring. Yet I believe that a court might appropriately hold that if a plaintiff presents a careful statistical analysis suggesting substantial discrimination and if the defendant by way of defense can do no more than suggest that the discriminatory effect in the plaintiff's model may be due to the chance effects of unmeasured variables, the plaintiff is entitled to a directed verdict—that is, to be declared the victor as a matter of law.

First, to show the possibility of randomness is not to demonstrate its existence, and to show that it is more likely than not that twenty independent investigations will yield one case in which a crucial association is by chance significant is not the same as showing that it is more likely than not that the one significant relationship found resulted from chance.

Second, if we are concerned that women are discriminated against in the workplace for reasons related to sex, we may not wish to compound the underrepresentation by the bad luck of the disfavored group. Whether the substantial underrepresentation of women in the firm we are investigating resulted from intent or chance, we may, for many of the same reasons we are opposed to sex discrimination, still wish to have the company increase the proportion of women in its workforce to about the proportion of qualified women in its applicant pool.

Third, for purposes of deterrence we may not want to allow a com-
pany with a disproportionately small number of women employees to appear untouched by the legal process. Other companies with a propensity to discriminate may thereby be emboldened to do so. Moreover, to allow the defense of randomness may mean that companies that are in fact discriminating escape by using it, thus doing injustice in particular cases and lowering specific as well as general deterrence.

This is not to say that there are no arguments in favor of allowing the randomness defense. Chief among these is the fact that we may think it unfair to stigmatize a company and penalize it through a back pay award when it did not engage in any activity that selected women out for disfavored treatment. Moreover, as Professor Rubinfeld points out, to allow innocent companies to be held responsible for discrimination may induce many companies that do not discriminate to engage in economically wasteful attempts to avoid the appearance of discrimination.

However one resolves this value conflict, we should not lose sight of the more general point: ways of thinking statistically with clear implications for social science—when we try twenty independent tests and observe one significant result at the .05 level we do not shout "Eureka"—are not by that token directly transplantable to the litigation setting. Statistical experts can and should point to the problem, but what to do about it is a question for legal policy makers. The law decides; statistical experts must conform—not vice versa.

Ultimately, Professor Rubinfeld and I share the same concern. We would both like to see courts make better use of the statistical evidence that is with increasing frequency presented to them. This requires not just more sensitive and sophisticated approaches to the statistical aspects of litigation research. It also—and more fundamentally—requires that lawyers and statistical experts educate each other. Statistical experts must learn how their evidence articulates with legal values and best fits into the structure of a case. Lawyers and judges must understand not the technicalities of statistical analyses, but the underlying logic of the descriptions and tests that statisticians offer them. Members of each community must, in short, learn what it is to think like a member of the other. It is to this end that Professor Rubinfeld and I have written.

45. See Rubinfeld, supra note 1, at 1064.