Statistical Science
1988, Vol. 3, No. 2, 149–195

Employment Discrimination and Statistical Science

Arthur P. Dempster

Abstract. Good practice of legal statistics depends on a foundation of good statistical science. Causal inference from statistical data depends both on understanding of substantive causal processes and adequate principles of statistical inference. The paper makes a case that Bayesian reasoning is needed for statistical studies of employment discrimination. A model based on Bayesian principles is developed in detail and is used to show that any statistical estimate of the effects of employment discrimination must be adjusted from sources of knowledge outside the statistician's data. Econometric analyses, which suggest otherwise, are analyzed and criticized.

Key words and phrases: Bayesian inference, causal analysis, employment discrimination, direct and reverse regression.

1. INTRODUCTION

Statistical science is aimed at understanding real world phenomena, whether natural or social, through a combination of empirical data analysis and statistical modeling. To approach scientific analysis this way is in the tradition of Jerzy Neyman's work, and hence the following study of roles for statistical arguments in evaluating regression-based claims of employment discrimination reflects many scientific principles that Neyman supported in his long and admirable career as an innovator and leader. In particular, I share his beliefs that probabilities ought ideally to be obtained from empirical frequencies, and that an important test of reported inferences is long run conformity to specified properties such as 0.05 levels or expectations.

I differ from Neyman in using the subjective interpretation of probability, and in giving Bayesian reasoning rather than sampling arguments a more central place in the practice of statistical inference and decision making. I see no contradiction between the subjective interpretation of probability and the desirability of objective evidential bases for assumed models that incorporate subjective probabilities as technical constructs. In my opinion, the major concern about objectivity arises not over the issue of subjective versus nonsubjective interpretations of probability, but rather over the empirical evidence and scientific judgment required of a model builder who seeks objectivity when selecting a particular probability model for a specific real world analysis. Because an essential feature of science is that findings and knowledge can be communicated, we need to make evidential bases and judgments explicit, and be prepared to argue their use in reasoned debate.

In recent years, law courts have struggled to define standards for the use of statistical evidence in legal proceedings. Many statisticians and econometricians have participated in the process by serving as expert witnesses in adversary proceedings. Recent papers by Meier (1986) and Fisher (1986) survey the current state of the art and discuss many problems of professional standards faced by statistical experts in the field. My concern here is less with legal processes *per se*, but rather with statistical reasoning processes that in my view should be pursued when analyzing the circumstances surrounding legal cases. My remarks reflect experience as an expert mainly confined to a type of employment discrimination case where a protected class of employees alleges systematic unfair withholding of rewards such as compensation and promotion. Similar model building ideas could be applied more broadly in the social sciences, most directly when the phenomena involve repeated decision making. DeGroot, Fienberg and Kadane (1986) contains reviews of broader types of statistical issues in legal settings.

A previous paper (Dempster, 1984) set forth a framework with three levels of information: facts about each employee known to the statistician, facts about the employee known to the employer/decision maker and an ideal complete set of facts about the

Arthur P. Dempster is Professor of Theoretical Statistics, Harvard University, Cambridge, Massachusetts 02138. This article is a revised version of the IMS Neyman Lecture presented at the Joint Statistical Meetings, Chicago, August 20, 1986.

employee. The framework was used to describe a collection of obstacles facing a statistician seeking to infer the presence or absence of discrimination, or the extent of discrimination, practiced by an employer against a legally protected class of employees such as women, blacks or older persons. I also described a controversy over which of two forms of regression analysis, called direct regression and reverse regression, is the correct form to apply to employment data when seeking to estimate a discrimination effect, and I suggested that my framework could be used to resolve the controversy. Pratt (1986) concluded a brief discussion of Dempster (1984) with the remark, "But we must keep striving toward sensible modes of using statistics in legal and public arenas—think of the alternatives!" The present analysis attempts constructive development of the ideas in the previous paper.

The goal of Section 2 is to define a mode of statistical thinking that I believe to be appropriate and sufficient to deal with the problem. Then, in Section 3, I take the modeling process back to first principles, and set out the reasoning behind the type of model which I advocate. Simple mathematical analysis of my model indicates that either direct or reverse regression gives unbiased estimates of discrimination effects if corresponding assumptions hold, but that in general neither is valid, because the correct analysis depends on the value of a parameter that captures an essential aspect of the real world, but whose value cannot be estimated from the statistician's data. Two scientific hurdles are raised by this situation. One involves teaching statisticians and others to see clearly and explicitly that statistical analysis on its own rarely offers complete solutions to externally specified problems. Instead, statistical analysis brings us to the point of seeing clearly what the gap in our knowledge is, so that we may then address the second hurdle, namely, the difficult task of looking outside the data for evidence bearing on the missing information. Section 3 also relates the new models to more traditional descriptions of econometric models as typified by Goldberger (1984), and argues that Goldberger's framework locks in arbitrary assumptions which prejudice his conclusions and lead him to propose an irrelevant test for the validity of reverse regression. Finally, in Section 4, I discuss a conceptual dilemma, which can scarcely be avoided, either by employers or regulators. The question is: what to do if economic efficiency and legal avoidance of discrimination collide? I call this the problem of judgmental discrimination. I suggest that the model of Section 3 provides a basis for analyzing the dilemma, and leads to policy attitudes consistent with fair and reasonable ways to handle the problem.

Causal thinking is intrinsic to the issues. Is discrimination *causing* reduced pay for certain employees? Are actions or decisions by an employer *causing* employees to suffer ill effects of discrimination? Statisticians and their clients clearly need shared working principles leading to agreed understanding of potential causal inferences. One set of such principles, ably reviewed by Holland (1986a), represents a major intellectual contribution of statistics to science. The idea is to identify randomization-based studies as situations where causal inferences may be soundly drawn, and to evaluate other attempts at causal inference by judging whether their nonrandomized circumstances adequately conform to critical assumptions which are transparently satisfied in randomized studies. Statisticians such as Pratt and Schlaifer (1984) or Holland (1986a) are skeptical that many claimed causal effects in econometric or other social science models can in fact meet the assumptions.

Because the data sets on which claims of employment discrimination are usually based are administrative records, and therefore about as far from randomized studies as one can get, it might seem that statisticians concerned about their reputations for scientific credibility would declare causal inference impossible or worse, and leave the matter there. Nevertheless, while I accept the cogency and importance of the negative arguments, I believe that we bear professional responsibility for carrying the discussion further, but along lines complementary to the main issues raised so far.

Holland (1986a) classified discussions of causation into those addressing the meaning of the term cause, those attempting to understand or establish causal mechanisms in relation to a specific class of phenomena and those trying to identify and measure causal effects in specific situations. Although only the third type is directly concerned with inferring causal effects, implied attitudes and understandings involving the other two types must underlie any specific causal analysis of statistical data. For my views on the operational meeting of causation, I refer to Dempster (1987). An essential point is that causal language implies the presence of some action mechanism operating in the real world and yielding consequences called causal effects. The second type of discussion picks up from here and asks, for example, in an employment discrimination case, what are the causal mechanisms operating to determine rewards of employment? My claim that the analysis of Section 3 below makes a contribution to the statistics of employment discrimination rests on the adoption of an explicit view of the basic mechanism of reward determination which is at best left implicit in traditional econometric models.

The distinction does not involve the discrimination mechanism itself, because it seems safe and reasonable to regard the discrimination causal effect as a quantity deliberately subtracted from what would otherwise be a fair and proper reward, on the basis of some irrelevant employee characteristic. The problem arises over specifying the causal mechanism or mechanisms that determine the fair and proper standard. As near as I can understand, the traditional econometric model conceives the fair reward as determined by a mathematical formula with added random disturbance, which in mechanistic terms involves the action of substituting in a formula, throwing dice and performing a computation. Even though such activities, excepting the dice throw, are essential, they offer in themselves no explanation of how fairness is achieved. I propose to add another ingredient to the description of the causal mechanism; namely, the conception that the employer/decision maker actively engages in decision making under uncertainty, so for each employee computes a posterior expected reward based on a defensible methodology for assessing likely contributions of the employee. I do not claim that such a mechanism is entirely realistic or factual, although I do believe that it adds essential explanatory power to the traditional model. I do claim, as set forth in Section 3, that the proposed mechanism has consequences for the way a statistician should build models for assessing potential discrimination effects, and thence for clarifying to the statistician what are the gaps in information which must be filled from sources external to the data if the desired inferences about discrimination are to be completed.

2. THE MEANING OF STATISTICAL MODELS

The use of causal language is not ordinarily a source of misunderstanding in scientific communication. The same is not at all true of probabilistic language, and hence I disagree with Cox (1986) on the relative importance of establishing philosophical underpinnings for causation and probability. Even in the causal arena, confusion over the roles of variation and uncertainty generally dominates difficulties over understanding the meaning of the causal mechanisms.

I believe that statisticians who make their peace with conceptions of probability in terms which exclude or ignore subjective interpretation are depriving themselves of structure needed for an explicit and fully comprehensible account of uncertain inference. Moreover, when mathematical models are used to represent deterministic phenomena, there is little disagreement about what they mean, but, when probabilistic models are used to reflect chance and uncertainty, disagreement usually lies close to the surface and is often a serious impediment to scientific discourse. I wish therefore to establish with care my terms of discourse.

Simple normal linear models are adequate to capture the essential types of variation, and not infrequently are adequate in practice to represent subjective uncertainty. Consider a model of the form

$$Y = G\alpha + X\beta + e, \tag{1}$$

where each term in the equation denotes an $n \times 1$ vector and the ith row in the vector equation may be written

$$Y_i = G_i\alpha + X_i\beta + e_i, \tag{2}$$

where we suppose that Y_i denotes the salary of the ith employee, G_i denotes the gender of the ith employee (1 for male and 0 for female) and X_i denotes a $1 \times k$ vector of k characteristics of the ith employee. Gender is used instead of, say, race or age, because it is convenient to work with a concrete hypothetical example which is a clear cut dichotomy. The elements $e_1, e_2, \ldots, e_n$ of e are conventionally said to be independent $N(0, \sigma^2)$ random variables, whereas α, β and σ are said to be parameters of the model.

There is both consensus and disagreement over different aspects of the interpretation of the equations (1) or (2). There are two major noncontroversial points. One is the purely mathematical content of the model, including the mathematical meaning of probability, which has shifted little since the measure-theoretic revolution altered mathematical statistics in the 1930s and 1940s. The other is that, quite apart from any probabilistic interpretation, the equations have a well understood deterministic meaning, whereby each of the n employees is thought to be associated with actual numerical values Y_i, G_i, X_i and e_i, which obey the linear relation specified in equation (2). When we attempt to define the real world interpretation of the probabilistic content of the model, however, the consensus among theoretical statisticians disappears. The two main contenders may be labeled (i) the chance mechanism meaning and (ii) the personal measure of uncertainty meaning. As elaborated below, I believe that (ii) is the only satisfying meaning. Interpretation (i) is valid and useful in restricted circumstances, but then is both subsumed under and clarified by (ii). As remarked by Savage (1967), "The foundational difficulties in the definition of personal probability are less than those of other attempts to define probability, and the truth behind other attempts to define probability is correctly expressible through the theory of personal probability." See also Savage (1962, pages 102–103).

To apply the normal linear model (1) in the chance mechanism mode we need to believe that $e_1, e_2, \ldots, e_n$ are generated by a random mechanism which operates independently of X and G. There is a powerful attraction to thinking this way, because (2) then becomes a complete explanation of how each Y_i is determined, to wit, one first operates the chance mechanism to find e_i, then determines X_i and G_i for the ith employee, hence Y_i is computed from (2). If an employer actually operates this way, then a finding of $\alpha \neq 0$ is interpretable as a finding of sex discrimination, because the chance mechanism is sex-blind by definition, whereas $X_i\beta$ represents objective job-related characteristics, hence the remaining term $G_i\alpha$ is purely determined by gender, and so is discriminatory.

Courts have been hearing arguments close to this from statistical and econometric experts for a decade or so. These arguments often lead to a curious sort of legal mischief, because a judge may rule that a plaintiff's fitted model of form (1) establishes a *prima facie* case for discrimination, thus putting the burden of proof on the other side to show the absence of discrimination. The rebuttal is most often attempted from a similar causal model, but adding a term, say $X'\beta'$, on the right side of (1) to represent unobserved characteristics of each employee. It is then argued, correctly, that a value of α estimated from a regression analysis omitting X' is quite likely to be badly biased. But such arguments are often rejected by courts because the rebutters cannot actually produce the proper regression analysis including X' (because X' is by definition unobservable!), so are reduced to indirect argument which is judged too weak to overturn the more concrete *prima facie* case. Another rebuttal strategy is to produce a large number of regressions with alternative choices of X, which may exhibit variations in both magnitude and sign of estimated sex coefficients, and so may confuse the issue.

All of these arguments, both for and against the *prima facie* case, depend in an essential way on the chance mechanism as a causal factor. I believe that a damaging confusion operates here between the use of chance mechanisms as analogies and the use of chance mechanisms as realities, the former arising when we say that uncertainties in some real world situation are *like* the uncertainties in a dice game. The purpose of the analogy is mental illumination. Actual chance mechanisms in real world employment contexts, if they exist at all, have nothing like the major role prescribed in widely applied statistical models. Making this negative argument to a judge is not easy, however, because the other side can point to a huge literature on causal modeling in econometrics and sociology which evidently rests on similar assumptions of randomness. I believe that much stochastic modeling in these disciplines is undermined by a dependence on fictitious chance mechanisms.

How should we proceed? One position is to reject all scientific use of probabilistic reasoning unless the real world basis of the chance mechanism is firmly established. Certainly this would rule out all use of statistical inference procedures on data from observational studies in the social and medical sciences, and might seriously challenge many uses of stochastic models in natural and engineering sciences contexts. I believe there is a better way, namely, to recognize that chance mechanisms are only one way to establish scientifically acceptable personal measures of uncertainty, thus turning attention to the difficult but necessary task of establishing broader standards.

I believe that reasonable probability models may be constructed both for situations where chance mechanisms are identifiable in the real world and for situations where no such mechanisms are implied. In fact, there is a critical common element of human judgment in both varieties. For example, in a coin tossing situation, where the concept of physical chance mechanism may be plausible, application to specific real world circumstances requires a judgment that the uncertainty representation is symmetric under all permutations of the order of the tosses. On the other hand, if we wish to learn from 100 past experiences of a type of situation, in order to quantify uncertainty about the next experience of the same type, then, in circumstances where no actual chance mechanism is remotely in view, we may often be willing to agree that prior uncertainty about the 101 experiences (i.e., uncertainty before experiencing any of them) should treat them symmetrically. The objectivity of probability models constructed for either variety of situation requires that our colleagues accept the corresponding symmetry judgments as reasonable, and either type of symmetry should be scrutinized in much the same way by trying to identify asymmetries in the real world repetitions involved. Jacob Bernoulli's study of the binomial distribution leading to the law of large numbers (Bernoulli, 1713) explicitly used the urn model as an abstract representation of the problem of learning from past experience. We should do the same, while explicitly recognizing what we are doing.

The attractive force that binds the two types of situation together is the interpretation of probability as a personal measure of uncertainty. For example, returning to model (1), the subjective interpretation is relevant to an individual analyst who knows the numerical values of selected X_i and expresses personal uncertainty about the unknown numerical values of the corresponding Y_i by describing them as independent with $N(\alpha + X_i\beta, \sigma^2)$ distributions. Specific numerical probabilities calculated from these distributions according to the usual rules of mathematical probability are specific personal measures of uncertainty. In the coin tossing example, the probabilities are relevant to the analyst's uncertainty about prospective outcomes. And so on, across all applications of formal probability models.

The historical record as chronicled by Hacking (1975) shows that computed probabilities were proposed for decision making in games of chance by distinguished scholars beginning in the mid 17th century, and that within decades the further proposal to extend such uses to wider scientific spheres was well established. In other words, probability is a basic quantitative technical construct like force or temperature that we learn to measure, to enter into computations and to apply in various classes of situations. Individual probabilities are meaningful because we have learned how to use them in appropriately circumscribed situations to represent our uncertainty about unknown factual circumstances.

We do need to put more effort into better understanding of what constitutes acceptable judgmental and evidential bases. Although the theory of probability is in a class by itself when it comes to guiding analysts to inferences and decisions under uncertainty, I do not agree with the views of some Bayesian advocates that subjective probability should be used universally, making little or no distinction between sources that are truly subjective in the ordinary as opposed to technical sense of the term, and sources that are shared among communicating analysts. Although caution is advisable, we should explore frameworks which weaken Bayesian technology in return for more acceptable accounts of model construction, while retaining subjective probability interpretation as central to the semantics. The theory of belief functions is an alternative I take seriously (Dempster, 1968; Shafer, 1976, 1987), but no detailed models are as yet in place for statistical practice, while Bayes is at last becoming practically feasible.

The subjective interpretation requires that every probability be understood as a conditional probability given the current state of knowledge of the interpreter. Often the intended current state is only implicit, but must be clear from the context of the discussion, as, for example, in the case of a chance mechanism, where the associated probabilities are understood to be conditioned on prospective unbiased operation of the mechanism. On the other hand, if the world under study is plainly not governed by chance mechanisms, then stochastic modelers need to think explicitly about whose uncertainty they are describing and under what conditioning. I often find probability modeling assumptions in econometric studies to be opaque because no attention is given to describing who is supposed to be subjectively interpreting the probabilities given what current state of knowledge, and hence no clear basis exists for how the models were constructed, by whom, or why.

3. AN ECONOMETRIC MODEL FOR EMPLOYMENT DISCRIMINATION

Stochastic models are needed in employment discrimination studies because the economic processes are not fully known or predictable to any actor or observer in the system. I assume the presence of a primary actor called the *analyst* or *statistician* whose tasks are first to formulate models representing his or her uncertain knowledge about relevant factual aspects of the system, and second to compute and report posterior probabilities and expectations given the available data, including the posterior distribution of the effects of discrimination in the system under study. It is understood that the models are constructed to be as objective as possible in the sense of laying out the evidence and arguments in favor of the proposed representations, so that coexperts and consumers alike may be encouraged to place credibility in the inferences. In the sequel, if an assumption is referred to without specific attribution, as though made by the author, the reader may assume that the author is being identified with the analyst/statistician, as a matter of convenience.

There is also a secondary actor in the formal system called the *employer*, who represents the decision maker or set of decision makers with responsibility for setting the reward of each employee. The basic unit of statistical analysis is a *decision* to supply a reward Y_i to a particular employee, leading to a sample of n such decisions and associated rewards $Y_1, Y_2, \ldots, Y_n$. The reason behind identifying the units with decisions rather than with employees, besides the fact that the data base may refer to several decisions on the same employee, is to emphasize that each unit is associated with the activation of a decision making mechanism. Specifically, it is assumed, as part of the analyst's model formulation, that *fairness* on the employer's part means that the employer determines $Y_1, Y_2, \ldots, Y_n$ by calculating n separate Bayesian posterior expectations of the true worth measures $Y_1^{**}, Y_2^{**}, \ldots, Y_n^{**}$, which are assumed to exist, but to have values unknown either to the employer or the analyst. Note the implication that the employer sets up n different probabilistic representations of uncertainty, one for each unknown Y_i^{**}.

The realism of the assumed Y_i^{**} may be questioned, for example, because it is impossible to disentangle the contributions of individuals from those of a group, or because the utility of an employee's contributions is just too hard to define and measure. It may be difficult to measure reward Y_i too, especially if intangibles are included. For purposes of the present theoretical discussion, I assume the existence of Y_i and Y_i^{**} as a minimal set of constructs necessary to get at a difference $Y_i^{**} - Y_i$ which is to be interpreted as a measure of employee deficit. Note the implicit assumption here that Y_i and Y_i^{**}, and Y_i^{*} to be defined later, are measurements on a scale such that sums and differences are meaningful.

As proposed in Dempster (1984), the statistician's model assumes that a vector X_i of employee characteristics associated with the ith decision is available both for purposes of model construction and of drawing final inferences. The statistician also assumes that the employer possesses a more extensive vector X_i^* of characteristics associated with the ith decision, and that both believe there is a still more comprehensive vector X_i^{**} of characteristics including the information needed to determine the true worth measure Y_i^{**}.

Although the basic units of analysis are called a sample of decisions, they are not assumed to be a random sample from any defined population. They are, however, assumed to be exchangeable in the sense of de Finetti, for purposes of constructing a probabilistic model. This device, which is surely the *sine qua non* of Bayesian model construction, simply means that the formulation must have mathematical symmetry under all permutations of the indices i, implying that differences among the circumstances surrounding individual decisions must be explicitly built into the subscripted scalars and vectors appearing in the formulation. Because the latter condition may be contradicted by experience, there can be nothing immutable about a particular assumption of exchangeability, but whatever degree of objectivity most statistical inferences based on samples has is dependent on a consensus about some exchangeability hypothesis.

The famous de Finetti theorem on exchangeable random variables permits us to think of the sample of decisions as the uncertainty equivalent of a multivariate random sample where the variables are the list of properties of each decision. Deciding to represent uncertainty as equivalent to a multivariate sampling hypothesis leads tò the further questions of what family of multivariate distributions to specify, and, for Bayesian statistical analysis, what prior distribution over the family to specify. The main tradition in applied econometrics is to specify sampling distributions crudely from first and second conditional moments whose values are established by a mixture of assumption and estimation from data. For purposes of Bayesian analysis, moment assumptions can be enhanced to normal models. The robustness of the resulting simple analyses in actual practice remains a tricky question which, from a Bayesian perspective, can only be addressed on a case by case basis by computing with more realistic models, something which is impossible given the current primitive state of the art. Because the present discussion refers only to a hypothetical situation, however, there is no question of faithfulness to an actual population distribution, and I assume simple normal models.

Another simplifying assumption made here about the hypothetical situation is that the statistician's sample size is large enough to render parameters such as α, β and σ^2 effectively known precisely. In practice, of course, sampling error is usually large enough to be interesting, and influential regarding model choice, but here the intention is to focus the main attention on bias effects which are unrelated to sample size. A similar assumption was made by Pratt and Schlaifer (1984).

In econometrics, a linear model such as (1) is often described as a causal model. One motivation for the term is simply that economic science is directed at elucidating the causes of economic phenomena, and econometric models are seen as contributing in some way to this goal. But what is the causal mechanism reflected in (1)? A possible causal interpretation of part of the model is to describe e as arising from the operation of a chance mechanism. I argued in Section 2 that belief in such mechanisms in the world of real employment decisions is questionable, and I suggested that the probability aspects of the model should be interpreted noncausally. Perhaps sensing the weakness of the chance mechanism view, Goldberger (1984, page 108) first described the term e in the model (w in his notation) as a "disturbance," and then went on to identify the "disturbance" as representing "additional information available to the employer but not to the statistician." Pratt and Schlaifer (1984, page 11) quote similar phrases from several well-known econometrics textbooks.

What causal factors should enter a causally interpretable stochastic econometric model of employment discrimination? Although causal analyses can be complicated endlessly by digging deeper and deeper, I believe we need only consider two primary factors, corresponding to control and treated in the language of experimental statistics, where the control causal mechanism defines a standard against which possible discrimination is to be determined. Specifically, the control mechanism which I propose for adoption by the statistician has the employer actively processing the information available on each employee (i.e., G and X^*), and computing

$$Y^* = E(Y^{**}|G, X^*). \tag{3}$$

The second causal mechanism is overt sex discrimination whereby the employer adds α' to Y^* to determine the pay of males while paying females only Y^*, i.e., the employer sets

$$Y = G\alpha' + Y^*. \tag{4}$$

Both of the causal mechanisms just described are implicit in standard econometric analyses. For example, Goldberger (1984) uses α as an "assessment of discrimination," where I use α with various superscripts for several different measures of discrimination. He uses p for the "employer's assessment of productivity," which I believe corresponds to my Y^*. He does not introduce Y^{**}, however, and so does not explicitly represent the mechanism (3). Absent the mechanism, in my opinion, the various scientific, ethical or legal interpretations of the model cannot be understood. In particular, the basic distinction between judgmental and prejudicial discrimination, which I introduce below and discuss further in Section 4, requires explicit representation of Y^{**}.

Next I explain how the model defined by (3) and (4) can provide a statistician with a framework to estimate the effects of employment discrimination. The precise details of the model are as follows. The characteristics of individual employees, other than gender, are treated as jointly normal within employees and independent across employees. The covariance matrix is taken to be the same for all employees, while there are two possibly different mean vectors, one for each gender group. Because all conditional distributions of one subset of characteristics given the values of another subset of characteristics are normal with mean linear in the conditioning set and residual independent of the conditioning set, it is possible to construct linear models galore. Although all such models have predictive interpretations, only those associated with identified causal mechanisms merit causal interpretation.

The translation of (3) into a linear model, namely,

$$Y^{**} = G\alpha'' + X^*B^* + e^{**}, \tag{3*}$$

does have an identified causal component, namely,

$$Y^* = G\alpha'' + X^*\beta^* \tag{5}$$

in (3*) reflecting the employer's conditional expectation judgment, whereas Gα' in (4) reflects overt sex discrimination.

The interpretation of e^{**} is that it consists of employee characteristics not available to the employer at the time of decision making, i.e., it could be expressible as $e^{**} = X^{**}\beta^{**}$. *A critical mathematical property of the model is that the conditioning operation* (3) *guarantees that e^{**} is independent of X^* and has the same mean for both gender groups* (*which can be taken to be* 0). Because in general there is no reason to expect employee characteristics to have the same means in both gender groups, and there is plenty of empirical evidence to the contrary, it is important to stress the special reason for the equal means property here, namely, that the causal component Y^* in (3*) is obtained from conditional expectation.

Combining (3*) and (4) into a single equation yields

$$Y = G\alpha^* + X^*\beta^*, \tag{6}$$

where

$$\alpha^* = \alpha' + \alpha''. \tag{7}$$

One consequence of (6) is that there is no way given Y, X^* and G, let alone given the more restricted actual data Y, X and G, to separate α^* into the two components α' and α''. I call α' the effect of prejudicial discrimination because it carries the overt discrimination represented in (4). I call the other component α'' *judgmental* discrimination because it is a technical form of gender effect created as a byproduct of the employer's use of a basic tool of decision analysis. The problems posed by the presence of judgmental discrimination are deferred to Section 4. Here, we proceed to the problem of how to estimate the total discrimination effect α^*.

The problem would of course be trivial if the available data were Y, X^* and G, instead of the actually available Y, X and G, i.e., we would simply solve the equations (6) and obtain values for α^* and β^*. What statistical experts actually do, literally thousands of times, is carry out regression analyses based on the model (1) which superficially resembles the postulated true mechanism (6). Can one relate the α from (1) with the α^* from (6)? If so, can one adjust the incorrect α to obtain α^*? The answer to the first question is yes, easily. Formulas for bias in estimated regression coefficients when the wrong X is used have a long history, and I derive simple expressions below for my model. The answer to the second question is no, not easily.

Surprisingly, many statisticians and econometricians appear untroubled by the fact that the available α and the true α^* are different and seem content to report α as a valid measure of discrimination. One possible reason is that the bias $\alpha - \alpha^*$ is judged small enough to ignore. A second reason may be confusion over the relative scientific merit of the alternative models. After all, model (1) can be derived rigorously from the normal assumptions, as a decomposition of Y whose terms describe probabilistic prediction of an uncertain Y given an observed X. Even though the valid interpretation of (1) is based only on statistical association and not on established cause, the term causal model is so often applied indiscriminately in econometrics that many practitioners may not distinguish between labeled causal and actual causal. A third reason, which can operate synergistically with the second, is a type of default logic. If a model specifies parameters which cannot be estimated from the data, or in econometric terms are not identifiable, there is a tendency to blame the model rather than the real world, and to choose the model whose parameters are identifiable. The first line of reasoning may sometimes be successfully applied on a case by case basis (cf.

Krasker and Pratt, 1984). I regard the other two arguments as specious. Causal interpretation is essential to inferring discrimination as a cause and cannot be taken lightly. There is no logical connection between whether or not a parameter exists and is important in the real world, and whether or not we are lucky enough to have data permitting estimation of the parameter. If the data are not there, the only course open to the objectively oriented Bayesian is to seek better data, or, much the same, to seek objective sources of prior distributions.

Despite these negative remarks, there are excellent reasons for beginning a study by fitting model (1) and estimating α. One reason is to find out what the data can legitimately tell us, namely, how well Y can be predicted from X. A second reason is more technical. It turns out to be easier to give interpretable mathematical expressions for the bias $\alpha - \alpha^*$ than for α^* directly. This is due to the simplicity of the process of adding variables to a regression analysis.

If we think hypothetically of carrying out the regression analysis indicated by (3*) in two stages, first regressing on X, and second bringing in the information in X^* not contained in X, we may write

$$X^*\beta^* = \tilde{X}_1 + \tilde{X}_2 \tag{8}$$

where $\tilde{X}_1$ has the form $X\beta'$, while $\tilde{X}_2$ has the form $X^*\beta''$ and the linear compounds represented by $\tilde{X}_2$ are uncorrelated with X. Substituting (8) into (6) leads to

$$Y = G\alpha^* + \tilde{X}_1 + \tilde{X}_2, \tag{9}$$

hence comparison with (1) shows that $\beta' = \beta$ so that

$$\tilde{X}_1 = X\beta, \tag{10}$$

and

$$G\alpha^* + \tilde{X}_2 = G\alpha + e. \tag{11}$$

If we denote the male and female population means of $\tilde{X}_2$ by μ_{M2} and μ_{F2}, and take population averages of (11) for $G = 1$ and $G = 0$, then, because e has zero means for both gender groups, we find that $\mu_{F2} = 0$ and $\alpha = \alpha^* + \mu_{M2}$, hence

$$\alpha = \alpha^* + (\mu_{M2} - \mu_{F2}). \tag{12}$$

In words, the bias from using α in place of α^* is given by the difference of male and female population means of the additional predictive variable $\tilde{X}_2$ known to the employer but not to the statistician. The reason for retaining μ_{F2} in (12) is that the condition $\mu_{F2} = 0$ is an artefact of the particular choice of gender indicator G. Formula (12) holds for any gender indicator such that the male-female difference has absolute value unity.

The statistician's data Y and X provide α and $\tilde{X}_1$, assuming effectively infinite sample size, but is devoid of information on $\tilde{X}_2$, and thence on μ_{M2} and μ_{F2}. It is easy to construct artificial X^* consistent with given data Y and X such that μ_{M2} and μ_{F2} have arbitrary values. Such arbitrary values may often be unreasonable on *a priori* grounds, but there is no magic cure to be found in the data for the bias in (12). In this sense, we know that it is hopeless to try to solve the problem by replacing the traditional "direct" regression based on (1) with an alternative form called "reverse" regression, whatever the definition of reverse regression. Still, the reverse regression story is fascinating, and leads me to conclude that reverse regression has a possible role in helping the statistician who is serious about developing a prior (= posterior) distribution for the bias $\alpha - \alpha^*$.

The original motivation for the statistical method called reverse regression, as well as for the contrasting terms direct and reverse, comes from contrasting definitions of "fairness" which are virtually free of stochastic or causal modeling assumptions. Both approaches agree that each employee should be paid exactly what he or she deserves, and then ask for a substitute principle to be used in the real world where such perfection is not achievable. In the first approach, the principle is to require that, given equal qualifications, males and females should be paid the same on average. Ordinary, or direct, regression based on (1) is seen as a means to obtain a suitable qualifications measure to be used as a practical standard for judging such equality of pay averages over gender groups. It follows that $\alpha = 0$ is the criterion for no discrimination, and α is the amount to be added to female pay to achieve parity with males. This simple line of reasoning sounds appealing, until it is realized that the choice of standard is far from innocuous. For example, I also support the principle, but with the more appropriate $X^*\beta^*$ used in place of $X\beta$. The second approach reverses the roles of pay and qualifications and suggests that the criterion for no discrimination should be that males and females with given pay should on average have equal qualifications measure. Statisticians contemplating reverse regression are also naturally drawn to $X\beta$ from (1) as a suitable composite qualifications measure, hence testing for discrimination requires looking at the "reverse" regression of $X\beta$ on Y and G. If the gender groups have parallel but different regression lines, as implied by the assumption of common covariance structures in the two gender groups, the remedy is to adjust each female's pay by the amount required to bring the regression lines into conformity. The problem is that the required shift is different in the case of reverse regression from the case of direct regression.

All statisticians know that the regression lines of Y on X and of X on Y have different slopes, but the apparently paradoxical consequences for multigroup data have provoked extended discussion only recently. The proposal to use reverse regression as a standard for employment discrimination was put forth by Birnbaum (1979). The idea was independently conceived by Harry Roberts in the summer of 1979, at a time when he was carrying out massive analyses of employee records at the Harris Bank in Chicago, in preparation for a hearing on charges that the bank was practicing discrimination. There has been a long subsequent literature, including Conway and Roberts (1983), with subsequent commentaries and rejoinder in the April 1984 issue of the *Journal of Business and Economic Statistics*, and Goldberger (1984). The key point is that direct and reverse regressions often give conflicting messages. In situations where males are on average more qualified than females, it often happens that the average male salary exceeds the average female salary among employees with a given qualifications measure, suggesting discrimination against females, whereas simultaneously among employees with a given salary the average qualification measure of males exceeds the average qualification measure of females, suggesting discrimination against males. The papers cited above contain many examples of the phenomenon derived from actual data. Simple tables of averages over qualifications variables within salary categories, when compared with averages over salaries within qualification categories, are sufficient to show striking results.

I have already argued above that neither method can be proved to give an unbiased assessment of discrimination when allowance is made for the difference between the employer's qualifications measure $X^*\beta^*$ and the statistician's constructed $X\beta$. I now derive an expression for the bias in the reverse regression measure to complement the expression (12) for the direct regression bias. In addition to the notation μ_{M2} and μ_{F2} for the male and female means of $\tilde{X}_2$, we need the notation μ_{M1} and μ_{F1} for the male and female means of $\tilde{X}_1$, and notation τ_1 and τ_2 for the variances of $\tilde{X}_1$ and $\tilde{X}_2$ which the simple model represents as common to both gender groups. To obtain equations for the regression lines of $\tilde{X}_1$ on Y, we may use (9) to express the regression coefficient of $\tilde{X}_1$ on Y as $\mathrm{COV}(\tilde{X}_1, Y)/\mathrm{VAR}(Y) = \tau_1(\tau_1 + \tau_2)^{-1}$. Thus, the equations of the male and female reverse regression lines are

$$(13) \quad x - \mu_{M1} = \tau_1(\tau_1 + \tau_2)^{-1}[y - \alpha^* - \mu_{M1} - \mu_{M2}]$$

and

$$(14) \quad x - \mu_{F1} = \tau_1(\tau_1 + \tau_2)^{-1}[y - \mu_{F1} - \mu_{F2}].$$

The reverse regression estimate of the discrimination effect in favor of males is the male-less-female distance between these lines measured on the y scale. Denoting this effect by α_R, it follows easily from (13) and (14) that

$$(15) \quad \alpha_R = \alpha^* + (\mu_{M2} - \mu_{F2}) - [\tau_2/\tau_1](\mu_{M1} - \mu_{F1}).$$

A comparison of (12) and (15) shows clearly how a change from the "direct" assessment α to the "reverse" assessment α_R can easily switch the sign of the effect.

Note that all the quantities appearing in the additional term $[\tau_2/\tau_1](\mu_{M1} - \mu_{F1})$ are determined by the statistician's data Y, X and G, so that the problem of assessing the bias in α_R is no different from the problem of assessing the bias in α, i.e., for a personalist it is still the problem of assessing a probability distribution for $(\mu_{M2} - \mu_{F2})$, given whatever evidence can be brought to bear on the question. I do think that some statisticians facing this admittedly formidable task would be helped by having the additional quantities associated with reverse regression in view. In particular, the condition that α_R is unbiased can be written

$$(16) \quad \mu_{M2} - \mu_{F2} = [\tau_2/\tau_1](\mu_{M1} - \mu_{F1}).$$

This says that an unknown gender difference is a specified multiple of a known gender difference. I might or might not be able to find evidence that the specified multiple is correct, but I believe I would often find it easier to think about what the correct multiple should be than to think about the more abstract $\mu_{M2} - \mu_{F2}$ directly.

Finally, I wish to reconcile my analysis of direct versus reverse regression with that of Goldberger (1984). Goldberger clearly sets out to refute "various claims" and "certain impressions" made and left by "critics of direct regression and proponents of reverse regression." His approach is to describe three possible models, each meant for "causal" interpretation and furnished accordingly with a "path diagram." Under Model A, "Multiple Causes" (Goldberger's quotation marks), the direct regression estimate of discrimination effect is shown to be unbiased. Under Model B, "Multiple Indicators," the reverse regression estimate is shown to be unbiased. Under Model C, "Errors in Variables," neither estimate is generally unbiased. Goldberger argues that Model B "is the only known specification under which reverse regression provides a valid estimator of α," where α denotes the effect of discrimination against females in all his models. He shows that Model B has a testable property, and gives an example where a significance test on the statistician's data rejects the testable property, and another example where it does not, perhaps in the latter case because the sample is too small. He concludes that "reverse regression should not be taken seriously unless accompanied by the information needed to test the restrictions of the multiple indicator model." He refutes certain claims for reverse regression made on the basis of Model C, but I did not find a clear discussion of how he proposes to deal with the bias of the direct regression estimator under Model C. The reader is left with little doubt that Goldberger considers direct regression to be the main contender.

I find Goldberger's analysis to be flawed in both general and specific ways. The general defect is the (for me) mysterious way in which "causal" models appear and depart without any detailed connection to causal processes in the real world. Goldberger (1984) misinterprets a remark of Dempster (1984): by being "somewhat skeptical about the existence of a chance mechanism," I was not arguing that an error term in a certain linear model has zero variance, but rather that the whole concept of error term is being inappropriately borrowed from stochastic models adopted in quite different fields of science. Goldberger (1984) strengthens my skepticism.

More specifically, Goldberger's Models A and B make strong assumptions which are essential to his claim that Model A justifies direct regression, and that Model B facilitates rejecting reverse regression through a significance test. When these assumptions are removed, both models are consistent with my causal model. If condition (16) is assumed, my model provides a counterexample to Goldberger's hypothesis that Model B is the only model supporting reverse regression. I demonstrate below that Goldberger's test can only invalidate unnecessary assumptions whose removal leads back to my model.

Consider Model A, "Multiple Causes," which I take to be Goldberger's model of choice. Model A is specified by three equations,

$$\text{“(30a, b, c)} \qquad y = p + \alpha z, \quad p = \beta' x + w, \qquad x = \mu z + u\text{ .”}$$

Equation (30a) is truly causal because it represents overt discrimination, and is the same as my (3) with the different symbols y for Y, p for Y^*, α for α'' and z for G. The second and third equations are formal linear models of a sort widely encountered in statistics, where the terms w and u are random "disturbance" terms assigned zero means, while $\beta' x$ and μz are systematic components.

Equation (30c) is easily understood in a descriptive sense as saying that the qualifications x (or X in my notation) differ in gender means. Because x is observed, this gender difference is an empirical fact in most applications. But the possible causal interpretations of (30c) are unspecified, and hence, in my view, the label causal should be avoided. For example, can it be meaningful to say that a gender difference in years of schooling is caused by gender? I do think that such terminology can convey scientific meaning as a kind of understood shorthand for causal mechanisms related to gender. It is no doubt true that complex social forces operate to create educational disadvantages to females in certain labor markets, and advantages to females in other markets, with a net bias toward helping males to obtain more prestigious and higher paying jobs. But all statistical variation, among individuals as well as among groups, is ultimately susceptible to such open-ended causal explanation. I suggest that it is preferable to maintain the term descriptive for variation only vaguely explained in causal terms, while saving the term causal for specified explanations.

Equation (30b) is the most interesting of the set, because it must carry what I called above the control causal mechanism. My version of (30b) is

$$(17) \qquad Y^* = G\alpha'' + \tilde{X}_1 + \tilde{X}_2,$$

which follows formally from (5) and (8), and originally from (3) which specifies the decision making mechanism of the employer. By definition, Y^* and p coincide, and my $\tilde{X}_1 = X\beta$ and his $\beta' x$ are the same. And, because w is described as "the additional information available to the employer but not to the statistician," it must be that w and $\tilde{X}_2$ are the same. The differences between (17) and (30b) are therefore that the former includes the term $G\alpha''$, which introduces judgmental discrimination, and that I allow for nonzero gender difference in $\tilde{X}_2$, whereas Goldberger arbitrarily assumes zero gender difference in w. The result is his conclusion that direct regression leads to unbiased estimates of discrimination effects, whereas my formula (12) leaves the statistician not knowing the bias. There is typically a substantial observed gender component in $\tilde{X}_1$; hence it is plausible that a similar component would be found in $\tilde{X}_2$ if it could be observed. I suggest that default logic is inappropriate when there is good reason to doubt the default option.

Consider next Goldberger's Model B, "Multiple Indicators," specified by

$$\text{“(40a, b, c)} \qquad y = p + \alpha z, \quad x = \gamma p + \varepsilon, \qquad p = \mu z + u.\text{”}$$

Again, the first equation represents overt discrimination, while the third equation makes descriptive sense, but lacks explicit causal explanation. Model B is unlike Model A in that the second equation also lacks causal sense, because the employer's assessment of productivity cannot be a causal determinant of job-related characteristics of the employee.

Goldberger motivates (40b) as reflecting a quote of Harry Roberts to the effect that job-related characteristics are "surrogates or proxies for productivity," which suggests to me that (40b) is meant for predictive interpretation. Earlier in his paper, on page 104, Goldberger notes "confusion" over an "elementary distinction" between "a proxy for p (in the imperfect correlate sense)" and "a fallible measure of p (in the strict errors-in-variable sense)." By omitting systematic effects other than p from (40b), he evidently opts for the fallible measures interpretation of Roberts's proxies.

Suppose we back up for a moment and ask what would have become of Model B if nonzero gender effects had been allowed in each of the equations (40b). Justification of reverse regression would not have followed. But the model would have been stochastically equivalent to my general model, which justifies neither direct nor reverse regression. Goldberger could then have sought the condition for reverse regression to be valid, which would have led to a single condition equivalent to (16), not to k conditions as in (40b). By arbitrarily assuming zero gender effects in all of the equations (40b), he introduces a strong mathematical assumption, namely, that the mean gender differences of each component of x should be proportional to the regression coefficients of these components on p, where the constant of proportionality is the mean gender difference on the employer's assessed productivity. The strong assumption of equality among k quantities is indeed testable, but failure of the proportionality does not invalidate (16), which is not testable because it contains terms which cannot be estimated.

In short, I believe that the models A and B of Goldberger (1984) can and should be reconciled with my model in such a way as to undermine his arguments for direct and against reverse regression. Both forms of regression are subject to bias that purely statistical methods are unable to correct.

4. JUDGMENTAL DISCRIMINATION AND OTHER PROBLEMS

If the statistician's problem is to learn as much as possible about the employer's reward setting process, then we need to address both the problem discussed in Section 3 of adjusting α^* for statistical bias, and the problem of separating α^* into its components defined in (7); namely, α' which I call the prejudicial discrimination effect because it is a simple, across-the-board, gender difference having no basis in productivity assessment and α'' which I call the judgmental discrimination effect because it arises in the course of a presumed honest attempt to assess productivity. The first problem could be solved by eliminating the information differential between a statistician's X and an employer's X^*, by some mixture of getting more information to the statistician or restricting the employer to more objectively measurable X^*. But, knowledge of α^* obtained this way would offer no help at all in decomposing α^* into $\alpha' + \alpha''$, which would require further empirical measures of $Y - Y^*$.

Is it important to separate α^* into its components? I believe it is scientifically important to do so, even if, as a matter of law or policy, α^* is defined to be the effect of discrimination. I believe also that, if legal experts or policy makers were to understand the technical aspects of judgmental discrimination that I will now attempt to explain, they might wish to work toward replacing α^* by a different standard. In Dempster (1984), I traced the idea of judgmental discrimination to Phelps (1972), who called it "statistical discrimination." The term "statistical discrimination" is a good one, but could be confusing in my context, because the "statistical" refers to statistical reasoning by the employer, whereas much of my paper focuses on a statistician engaged in a very different sort of statistical reasoning from that of the employer. Whereas Phelps (1972) stated that judgmental discrimination is "damaging" and "important for social policy to counter," Dempster (1984) suggested that the question needed "further elucidation and debate," which I now attempt.

The expectation operation in equation (3) expresses an essential principle governing applications of the theory of personal probability. There are two basic ways of getting into trouble in the long run if the principle is consistently violated. Both are consequences of the property that, if the principle is followed, then in a long sequence of trials there is no way to select on the basis of X_i^* and G_i at each trial a subsequence whose average differs from that of averaging the predictions of the expectation operation. It follows that, if in a long sequence of trials an employer consistently pays women an amount α'' more than indicated by (3), while paying men exactly the amount indicated by (3), then in the long run women will on average receive α'' more than their true Y^{**}, while men on average will receive exactly their true Y^{**}. This will happen within any qualifications class, and will surely make a *prima facie* case for discrimination against men (assuming positive α''). The second difficulty with using α'' to adjust female salaries for judgmental discrimination is that a competitor can select against the reward setting mechanism so that in the long run the employer will be left with only female employees paid on average an amount α'' more than their actual average productivity, while the all-male competitor has the advantage of paying exactly the right amount. These hypothetical arguments are based on highly idealized circumstances, of course, but the phenomena are surely real. It follows that adjustments for judgmental discrimination require regulations which operate successfully against market forces.

A different sort of problem with abandoning the principle of equation (3) is the lack of any principle to put in its place. For example, it might be thought that replacing α'' by zero in equation (5) would define an excellent nondiscriminatory reward setting mechanism. Such mechanisms can only lead to endless controversy, however, over what variables can legitimately be included in X^*, because some variables would be criticized as intrinsically tainted carriers of discrimination. I argued in Dempster (1984) that many variables could legitimately be regarded as so affected, in varying degrees, thus posing difficult problems of judgment for a regulator. Notice, however, that, if the definition of fairness implicit in equation (3) is maintained, there is no restriction at all on the variables admitted to X^*.

There are ways to address real problems of discrimination without violating the theory of probability. For example, if the problem is to redress the negative effects of past overt discrimination on the qualifications of a protected class, then the cost of doing so could be assessed as a tax on employers. Assuming that reasonable equity among competing employers can be maintained, then overpaying members of a disadvantaged class while they improve qualifications is certainly a reasonable form of tax. Various equivalent forms, both voluntary and enforced, are easily conceived. Affirmative action programs can be rationalized in this way.

Another way to proceed is to exploit subjective uncertainty in the use of the theory of personal probability. An employer's assessment of the expected productivity of employees in a protected class may be wrong in a way that social experimentation could effectively demonstrate, if such experimentation were feasible, which it rarely is due to time lags and prohibitive expense. The regulator may argue that an employer's productivity assessment based on past experience is invalid if it does not incorporate certain mechanisms, such as, for example, the effect of higher rewards themselves on the productivity of employees in a disadvantaged class. If there is an honest difference of personal probabilities, each backed by an explicit and credible analysis, then political forces may permit the regulator to prevail, no doubt sometimes correctly, and sometimes wrongly, as can be demonstrated only in the long run. Effective regulation of this sort obviously requires skill in causal modeling of the social phenomena concerned. Again, I believe that existing affirmative action agreements could be interpreted as legitimate guesses at fair procedures under assumptions about causal processes not yet formally expressed as computations.

Finally, a brief comment on technical statistics, specifically on implications of my modeling strategy for statistical analysis. An important point is that the principle (3) is in no way tied to normal models, but it does imply that modeling needs to be carried out on an additively meaningful reward scale. Another point is that equation (4) is too simplistic for most applications in that the effects of overt discrimination are unlikely to be equal across a protected class. Statistical modeling thus needs to become increasing sophisticated about coping with nonnormal, including nonlognormal, distributions, and in allowing for interaction effects. These complications can only increase the burdens of expressing prior uncertainties about quantities not econometrically identified by the data. Formal education of applied statisticians and econometricians needs reconstruction along lines which include uncertainty assessment of gaps routinely left by standard statistical analyses.

ACKNOWLEDGMENTS

I thank the Institute of Mathematical Statistics for the opportunity to present the 1986 Neyman Lecture, on which the paper is based. Comments by John Carlin and Stephen Stigler on earlier drafts have helped to improve the presentation. The work was facilitated in part by National Science Foundation Grant DMS-85-04332.

REFERENCES

BERNOULLI, J. (1713). *Ars Conjectandi.*

BIRNBAUM, M. H. (1979). Procedures for the detection and correction of salary inequities. In *Salary Inequities* (T. R. Pezzullo and B. E. Brittingham, eds.). Lexington Books.

CONWAY, D. A. and ROBERTS, H. V. (1983). Reverse regression, fairness and employment discrimination. *J. Bus. Econ. Statist.* **1** 75–85.

COX, D. R. (1986). Comment on "Statistics and causal inference," by P. W. Holland. *J. Amer. Statist. Assoc.* **81** 963–964.

DEGROOT, M. H., FIENBERG, S. E. and KADANE, J. B., eds. (1986). *Statistics and the Law.* Wiley, New York.

DEMPSTER, A. P. (1968). A generalization of Bayesian inference (with discussion). *J. Roy. Statist. Soc. Ser. B* **30** 205–247.

DEMPSTER, A. P. (1984). Alternative models for inferring employment discrimination from statistical data. In *W. G. Cochran's Impact on Statistics* (P. S. R. S. Rao and J. Sedransk, eds.) 309–330. Wiley, New York.

DEMPSTER, A. P. (1987). Causality, uncertainty, and statistics. Paper presented at the Symposium in Honor of I. J. Good, Dept. Statistics, Virginia Polytechnic Institute and State Univ., Blacksburg, Va.

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

GOLDBERGER, A. S. (1984). Reverse regression and salary discrimination. *J. Human Resources* **19** 293–318.

HACKING, I. (1975). *The Emergence of Probability.* Cambridge Univ. Press, Cambridge.

HOLLAND, P. W. (1986a). Statistics and causal inference. *J. Amer. Statist. Assoc.* **81** 945–960.

KRASKER, W. S. and PRATT, J. W. (1984). Bounding the effects of proxy variables on regression coefficients. Graduate School of Business Administration, Harvard Univ.

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

Phelps, E. S. (1972). The statistical theory of racism and sexism. *Amer. Econ. Rev.* **62** 59–61.

Pratt, J. W. (1986). Review of "W. G. Cochran's Impact on Statistics." *J. Amer. Statist. Assoc.* **81** 565–566.

Pratt, J. W. and Schlaifer, R. (1984). On the nature and discovery of structure. *J. Amer. Statist. Assoc.* **79** 9–21.

Savage, L. J. (1962). *The Foundations of Statistical Inference.* Methuen, London.

Savage, L. J. (1967). Implications of personal probability for induction. *J. Phil.* **64** 593–607. Reprinted in *The Writings of Leonard Jimmie Savage—A Memorial Selection*, published by ASA and IMS.

Shafer, G. (1976). *A Mathematical Theory of Evidence.* Princeton Univ. Press, Princeton, N. J.

Shafer, G. (1987). Probability judgment in artificial intelligence and expert systems. *Statist. Sci.* **2** 3–16.

Comment

Franklin M. Fisher

Arthur Dempster's paper has a good deal to say about the interpretation of probability models and causal thinking, much of it uncontroversial. Rather than discuss such matters in the abstract, however, let's consider the example of employment discrimination that Dempster uses and see what it is that he is really saying.

This is not hard for me to do, because I have encountered Dempster's views on previous occasions. I was a witness for the plaintiff in two employment discrimination cases, *OFCCP v. Harris Trust and Savings Bank* (Department of Labor Case No. 78-OFCCP-2) and *Cynthia Baran v. The Register Publishing Company* (Civil N. 75-272, U. S. District of Conn.). In both cases, I testified on matters of econometric and statistical principle rather than putting forward a study of my own, and Dempster testified for the defendant. This paper is largely based on my experience and testimony in those cases. (I believe—but do not know for sure—that, just as my own experience in employment discrimination cases has been as an expert assisting plaintiffs' counsel, Dempster's experience, to which he refers, has been as an expert assisting counsel for defendants.)

A particular employer is accused of sex discrimination. (As does Dempster, I take this as a leading example.) In general, this means that salaries paid to female employees average less than those paid to male employees. One possible reason for this discrepancy is discrimination; another is that male employees are more productive than female ones.

To examine the question of whether there is a gender-based wage difference holding productivity constant, a statistician estimates the model

$$Y = G\alpha + X\beta + e, \tag{1}$$

where (letting i denote values for a particular employee), Y_i denotes salary, G_i is 0 for female and 1 for male employees, X_i is a vector of observed employee characteristics (education, experience, age, etc.), and the e_i are assumed to be random variables (usually taken to be independent $N(0, \sigma^2)$, although this will play no role in the present paper). α, β and σ are parameters to be estimated, and it will aid discussion to assume that the sample size is sufficiently large to enable us to take such parameters as known with certainty. A positive value of α is taken to be evidence of discrimination against females.

What is wrong with such a procedure? Dempster points out several possibilities. In the first place, he suggests interpreting the stochastic element involved by assuming that the nondiscriminatory employer is computing

$$Y^* = E(Y^{**} \mid G, X^*), \tag{2}$$

where X_i^* is a vector of employee characteristics known to the employer (but possibly not to the analyst), Y_i^{**} denotes "true" employee productivity and Y_i^* denotes employee productivity as estimated by the employer in (2). Both Y_i^{**} and Y_i^* are assumed measured in monetary units to be comparable to wages, Y_i. Discrimination is to be interpreted as paying males more than Y_i^*, i.e.,

$$Y = G\alpha' + Y^*, \tag{3}$$

with $\alpha' > 0$.

This is not the only form that discrimination can take. Depending on the state of the outside labor market, discrimination is more likely to consist of paying females *less* than the employer truly thinks

Franklin M. Fisher is Professor of Economics, Department of Economics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.