URN MODELS FOR REGRESSION ANALYSIS, WITH APPLICATIONS TO EMPLOYMENT DISCRIMINATION STUDIES

BRUCE LEVIN* AND HERBERT ROBBINS†

I

INTRODUCTION

Randomization models can be used to test the statistical significance of differences between two groups, even in observational studies where there is concomitant information, with a meaningful and valid probabilistic interpretation. This article discusses one procedure for incorporating concomitant information into a randomization model, obtaining an urn model for regression analysis. It is motivated in part by the frequency with which standard regression models are used improperly to deal with questions of disparate treatment of sex or race in employment.

Although urn models have a very wide applicability, this article presents them specifically for the statistical study of wage disparity between men and women in sex discrimination litigation, an area of application that highlights the problems arising with conventional population sampling models in nonexperimental settings. A characteristic problem is the drawing of inferences in the absence of traditional safeguards found in the natural sciences: experimental control of causal factors, replication of observations, and randomization of experimental units across treatments or conditions. Expert statistical witnesses, who must present their findings without such luxuries and in the face of adversarial criticism, need to base their analyses on a minimum of assumptions in order to support their conclusions. The urn model approach requires much less in the way of assumptions than the population sampling approach and is especially attractive for this reason.

The key idea, with which this article begins, is to separate the mean wage disparity between men and women into an *explained portion* (with respect to a specific adjustment procedure) and an *unexplained portion* that reflects any departure from equal earnings for men and women after adjustment for concomitant information. The statistical significance of the unexplained portion is then assessed with a randomization model. The method may be regarded as a generalization of some conventional testing methods, such as the Mantel-Haenszel procedure[1] for

* Division of Biostatistics, Columbia University School of Public Health, the G. Sergievsky Center, and the N.Y. State Psychiatric Institute.

† Department of Statistics, Columbia University.

[1] J. FLEISS, STATISTICAL METHODS FOR RATES AND PROPORTIONS 173 (1981). The technique

dichotomous variables, to the case of continuous outcomes with regression adjustment for covariates. The procedure is described here in successive stages of elaboration.[2] The unexplained portion of the mean wage disparity emerges as a new measure of wage discrimination, and its relation to other measures, especially the sex-coefficient, is then clarified.[3] The article concludes with a discussion and criticism of conventional regression models in observational studies.[4] It suggests that urn models be given serious consideration as an alternative to standard multiple regression formulations in the absence of validated causal models of salary discrimination.

The emphasis here is on significance testing in a two-group comparison, rather than on the estimation of coefficients in a linear regression model. This emphasis is in line with the noncausal approach adopted here in which the unexplained portion of the mean wage disparity is used to test the null hypothesis that sex has no direct influence on salary. The unexplained portion is found to be not statistically significant, in the urn model sense made precise below, if a concrete, sex-neutral, chance mechanism would produce the observed disparity reasonably often. With such a finding there may be no need for further modeling. If the null hypothesis is rejected then further investigation into validating specific parametric models and estimating their parameters may be in order. The focus here is on the first step only. The difficulties involved in devising and supporting realistic causal models are well known and will not be discussed in this article. In particular, no consideration is given to the problems of model specification or bias in misspecified models.[5] It is assumed that the set of explanatory variables to be used for regression adjustment has already been determined. This article now turns to the definition of the unexplained portion of the mean wage disparity and the evaluation of its statistical significance.

II

URN MODELS AND REGRESSION

The analysis begins with the simple case of a single stratum of n individuals, homogeneous except for sex. This case is referred to as the *single stratum, no regression* case. Suppose there are m men and $w = n - m$ women, each with a known salary Y_i $(i=1, \ldots, n)$. The *mean wage disparity* between men and women is defined as the difference in sex-specific salary means,

$$d=1/m \ (\sum_i{}^{1} Y_i) \ - \ 1/w \ (\sum_i{}^{0} Y_i),$$

where Σ^1 indicates summation over men and Σ^0 over women. Under the hypothesis of no influence of sex on salary, the observed association of sex with salary is

derives its name from its intial documentation in literature by Mantel & Haenszel, *Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease,* 22 J. NAT'L CANCER INST. 719 (1959).

[2] *See infra* section II.

[3] *See infra* sections III and IV.

[4] *See infra* section V.

[5] For a note on the underadjustment phenomenon and related literature, see Robbins & Levin, *A Note on the "Underadjustment Phenomenon,"* 1 STATISTICS & PROBABILITY LETTERS 137 (1983).

just one of $\binom{n}{m} = \binom{n}{w}$ equally likely possibilities for the given numbers m, w, and $Y_1, \ldots, Y_n$. Thus, one may envision placing the numbers $Y_1, \ldots, Y_n$ on n chips in an urn, selecting m chips at random without replacement, and computing the difference, say D, between the average values of the chips withdrawn and those left in the urn. D simulates the disparity that would arise from the action of sex-neutral chance factors, and the random variable D is referred to as a *simulated mean wage disparity*. If the observed d lies well within the central portion of the sampling distribution of D, then it might reasonably have arisen by pure chance without sex discrimination. If, on the other hand, the observed d lies far out in one or the other of the tails of the sampling distribution of D, then the hypothesis of no influence of sex on salary is implausible.

The probability distribution of the simulated mean wage disparity D is, in principle, obtainable exactly. It can be calculated explicitly for small values of m or w and simulated by computer in all cases. This article considers only the large sample normal approximation to the distribution of D, but the adequacy of this normal approximation can and should be checked by simulation. Physical simulation may also be instructive in explaining the basis of inference to a court, and it serves to fix ideas about the meaning of the urn model. (One might even imagine deciding the question on the basis of whether a single simulated $|D|$ exceeds the observed $|d|$, although it is less arbitrary to rely on the probability distribution of D than to rely on a lottery.) One should examine the entire distribution of D, rather than just the usual critical points for hypothesis testing, since in this way one can evaluate not just the nominal significance level of d but also the likelihood of what a court may consider to be substantial differences. The entire distribution will be of special importance in cases when the exact or simulated distribution is markedly skewed or otherwise nonnormal.

The exact mean and variance of the simulated mean wage disparity D, under the hypothesis of no influence of sex on salary, can be shown to be

$$E(D) = 0 \qquad \textit{and} \qquad Var(D) = s^2 \cdot \frac{n}{mw},$$

where by definition

$$s^2 = \begin{cases} \frac{1}{n-1} \sum_{j=1}^{n} (Y_j - \mu)^2 & \textit{if } n > 1 \\ 0 & \textit{if } n \leqq 1 \end{cases}$$

and

$$\mu = \frac{1}{n} \sum_{j=1}^{n} Y_j .$$

The central limit theorem[6] for sampling without replacement from a finite population asserts that the random variable

[6] E. LEHMANN, NONPARAMETRICS: STATISTICAL METHODS BASED ON RANKS 352 (1975). *See also* S. WILKS, MATHEMATICAL STATISTICS 464 (1962).

$$Z = D \,/\, (s^2 \frac{n}{mw})^{1/2}$$

is approximately standard normal (mean 0 and variance 1) for large *m* and *w*, so that in practice we need only calculate the observed ratio

$$z = d \,/\, (s^2 \frac{n}{mw})^{1/2}$$

and refer it to the standard normal distribution to find the nominal level of significance. (Later, *z* is compared with the usual two-sample *t* statistic.[7])

The analysis next turns to the more common *several strata, no regression* case. The strata are assumed to be well-defined subgroups of the entire group (for example, subgroups formed by a factorial design of qualitative variables, as in the analysis of variance) and should have substantive meaning as a basis for statistical adjustment. An example would be stratification by department in a university. When there are *k* strata, let *i* run over the integers from 1 to *k* and define

$$m_i = \textit{number of men in stratum } i$$
$$w_i = \textit{number of women in stratum } i$$
$$n_i = m_i + w_i = \textit{number of people in stratum } i,$$

and

$$m = \Sigma^k_1 m_i = \textit{total number of men}$$
$$w = \Sigma^k_1 w_i = \textit{total number of women}$$
$$n = \Sigma^k_1 n_i = \textit{total number of people.}$$

Denote the salary of the j^{th} individual in the i^{th} stratum by Y_{ij}. The observed mean wage disparity *d* between men and women is

$$d = (\frac{1}{m} \sum_i \sum_j{}' \; Y_{ij}) - (\frac{1}{w} \sum_i \sum_j{}^{o} \; Y_{ij}) .$$

The remainder of this section proposes a decomposition of *d* into two parts:

$$d = \textit{explained portion} + \textit{unexplained portion.}$$

As opposed to the analysis of variance, this decomposition partitions the mean wage disparity, and not a sum of squared deviations. A number $\Delta\bar{r}$, introduced below, will represent the unexplained portion of the decomposition.

Before giving a general statement of the method of decomposition, the university example will repay some further consideration. Suppose that departmental affiliation is the sole explanatory variable under consideration. A positive observed mean wage disparity *d* would be explained, at least in part, if women tend to be found in departments with generally lower salaries, a possibility that is not in itself discriminatory (assuming no selective shunting of females into lower paid departments). A reasonable way to take this factor into account and separate *d* into explained and unexplained portions is the following. Suppose that each person's salary is placed on a chip (without any other information, such as sex) and *k* different urns are set up to contain the chips belonging to each department. Now if one randomly withdraws as many chips from each urn as there are men in

[7] *See infra* section IV.

the corresponding department, one will arrive at a random separation of employees' salaries into two groups—the chips withdrawn and the chips remaining in the urns—that simulates exactly the kind of disparity that would arise from the action of sex-neutral chance factors after adjustment for department. Again, one calculates the mean wage for the chips withdrawn and those left, takes the difference, and calls the resulting random variable D the simulated mean wage disparity. The expected value of D (averaged over all possible outcomes of the simulation) may be termed the *explained portion* of the observed mean wage disparity; it measures the average simulated mean wage disparity that would result solely from the actual differential representation of men and women in the various departments. The *unexplained portion* of the observed mean wage disparity d is defined to be the difference $\Delta\bar{r} = d - E(D)$. To assess the significance of this unexplained portion one asks how often $|D - E(D)|$ would exceed the observed $|\Delta\bar{r}|$. If the answer is "often," say in at least 5% of all simulations, then one is in the domain of purely random disparity. If the answer is "seldom," say in less than 2% of all simulations, then one has an indication of nonrandom disparity. Thus, the p-value derived gives the proportion of simulations in which the observed value of $|\Delta\bar{r}| = |d - E(D)|$ would be exceeded on the basis of a simple and clearly defined discrimination-free chance mechanism.

In general one sets up k urns exactly as in the example and considers the distribution of the simulated mean wage disparity D under the chance mechanism of stratified random sampling from the k urns, where D is the difference between the mean scores obtained by combining the k sets of m_i chips selected from each urn and the k sets of w_i chips remaining in the urns. This provides the distribution of D under the null hypothesis of no influence of sex on salary within each stratum. The expected value and variance of the random variable D are given by the formulas

$$E(D) = \sum_{i=1}^{k} \left(\frac{m_i}{m} - \frac{w_i}{w}\right) \mu_i$$

and

$$Var(D) = \left(\sum_{i=1}^{k} s_i^2 \frac{m_i w_i}{n_i}\right) \left(\frac{1}{m} + \frac{1}{w}\right)^2 ,$$

where

$$s_i^2 = \begin{cases} \frac{1}{n_i - 1} \sum_{j=1}^{n} (Y_{ij} - \mu)^2 & if\, n_i > 1 \\ 0 & if\, n_i \leqq 1 \end{cases}$$

and

$$\mu_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$$

is the mean salary of all n_i people in the i^{th} stratum $(i=1, \ldots, k)$. Under conditions similar to those for the single stratum case, or under the alternative assumption that k is large even if each n_i is small, the ratio $Z=(D - E(D))/Var^{1/2}(D)$ has approximately the standard normal distribution, and in practice one can refer

$$z = \frac{d - \sum_{1}^{k} \left(\frac{m_i}{m} - \frac{w_i}{w}\right) \mu_i}{\left(\sum_{1}^{k} s_i^2 \frac{m_i w_i}{n_i}\right)^{1/2} \left(\frac{1}{m} + \frac{1}{w}\right)} = \frac{\Delta \bar{r}}{\left(\sum_{1}^{k} s_i^2 \frac{m_i w_i}{n_i}\right)^{1/2} \left(\frac{1}{m} + \frac{1}{w}\right)} \qquad (1)$$

to the standard normal distribution to obtain nominal significance levels.

The decomposition of the mean wage disparity d into explained and unexplained portions is

$$d = E(D) + (d - E(D)) = E(D) + \Delta \bar{r},$$

and this article proposes $\Delta \bar{r} = d - E(D)$ as a measure of discrimination in the several strata, no regression case. An algebraically equivalent but perhaps more suggestive formulation of the urn model is the following. Rather than recording Y_{ij} on each chip, one records only the "residual"

$$r_{ij} = Y_{ij} - \mu_i \; (j=1, \ldots, n_i; \; i=1, \ldots, k).$$

For each individual employee we regard μ_i as the explained portion and r_{ij} as the unexplained portion of his or her salary. Then the explained portion of the mean wage difference, E(D), is simply the difference between the male and female averages of the explained portion of salary, while the unexplained portion of the mean wage difference, $\Delta \bar{r}$, is the difference between the male and female averages of the unexplained portions of salary. Stratified random sampling of m chips from the k urns then leads to a simulation of $\Delta \bar{r}$. This formulation suggests the following extension of the definitions to the *regression* case.

Consider again the case of a single stratum, but with various explanatory variables $X_j^{(1)}, \ldots, X_j^{(p)}$ *excluding sex* for employee $j=1, \ldots, n$. This section supposes that an *adjustment formula* $f(X^{(1)}, \ldots, X^{(p)})$ is given such that with it one can define

$$\textit{explained portion of salary for employee } j = \hat{Y}_j = \mu + f(X_j^{(1)}, \ldots, X_j^{(p)}).$$

The formula f is used to adjust the explained portion of salary, $\hat{Y}_j$, above or below the average salary μ, on the basis of the covariates $X^{(1)}, \ldots, X^{(p)}$, in the same way for men and women. The unexplained portion of salary for employee j is then defined as the residual,

$$r_j = Y_j - \hat{Y}_j = Y_j - (\mu + f(X_j^{(1)}, \ldots, X_j^{(p)})).$$

These definitions are extended in the obvious way to express the mean wage disparity as the sum of an explained portion plus an unexplained portion:

$$\textit{explained portion of mean wage disparity} = \left(\frac{\Sigma' \hat{Y}_j}{m}\right) - \left(\frac{\Sigma'' \hat{Y}_j}{w}\right),$$

and

$$\textit{unexplained portion of mean wage disparity} = d - \textit{explained portion} =$$

$$\Delta\bar{r} = \left(\frac{\Sigma' r_j}{m}\right) - \left(\frac{\Sigma^0 r_j}{w}\right).$$

The adjustment formula f is arbitrary up to this point, and this article envisions the use of whatever appropriate adjustments the case may warrant. For example, in a study of salary increases, an employer may use certain guideline formulas (generally nonlinear and time inhomogeneous) for calculating yearly standard increases, departing from the formula in special cases ("out of guideline"). The logical procedure would then be to examine the disparities in increases remaining after adjustment for guideline increases by the actual formulas. Assuming that the formulas themselves are fair and do not have a disparate impact on men and women, which is a separate question, the unexplained portion of the mean disparity in wage increases speaks directly to the impartiality of the out of guideline increases. Coupling this analysis with a study of initial placement and initial salary may provide a very convincing explanation of existing salary disparities. In this example there is no need to estimate any coefficients at all, since the actual guideline formulas are assumed to be available.

In the absence of such information, the adjustment formula could be the usual linear least squares function

$$f(X^{(1)}, \ldots, X^{(p)}) = b^{(1)}(X^{(1)} - \bar{X}^{(1)}) + \ldots + b^{(p)}(X^{(p)} - \bar{X}^{(p)})$$

where $b^{(1)}, \ldots, b^{(p)}$ are the ordinary least squares multiple regression coefficients. For purposes of comparing $\Delta\bar{r}$ with the more familiar sex-coefficient,[8] we shall henceforth assume this version of the adjustment formula. Note that for general f the residuals r_j may have a nonzero mean $\bar{r}$, although of course $\bar{r} = 0$ for the linear least squares adjustment formula.

The measure $\Delta\bar{r}$ represents that part of the mean wage disparity left unexplained after regression adjustment for concomitant variables $X^{(1)}, \ldots, X^{(p)}$. To assess the significance of $\Delta\bar{r}$, consider the urn model for the random selection of m residuals, and compute the ratio

$$z = \Delta\bar{r} \,/\, s\left(\frac{n}{mw}\right)^{1/2},$$

where

$$s^2 = \begin{cases} \dfrac{1}{n-1}\displaystyle\sum_{1}^{n} (r_i - \bar{r})^2 & n > 1 \\ 0 & n \leqq 1. \end{cases}$$

The final elaboration considers the *stratified regression* case. For stratum $i = 1, \ldots, k$ one assumes as given an adjustment formula f_i based on explanatory variables $(X_{ij}^{(1)}, \ldots, X_{ij}^{(p)})$ *excluding sex* for person j $(j = 1, \ldots, n_i)$. The explained portion of salary for person j in stratum i is now

[8] *See also infra* section IV.

$$\hat{Y}_{ij} = \mu_i + f_i(X^{(1)}_{ij}, \ldots, X^{(p)}_{ij}) = \mu_i + b^{(1)}_i(X^{(1)}_{ij} - \bar{X}^{(1)}_i) + \ldots + b^{(p)}_i(X^{(p)}_{ij} - \bar{X}^{(p)}_i),$$

where $bi^{(1)}, \ldots, bi^{(p)}$ are the stratum-specific least squares regression coefficients. One does not need to assume parallel surfaces across strata, although if the strata are too small one may wish to place restrictions on the coefficients to ensure nonsingularity. In special circumstances one may use the same regression coefficients across strata if there are compelling substantive reasons for doing so.

The residual $r_{ij} = Y_{ij} - \hat{Y}_{ij}$ represents the excess of a person's salary over his or her stratum-specific adjusted mean salary, and is the unexplained portion of salary for that person. These definitions are extended as before to the

explained portion of the mean wage disparity =

$$\left(\frac{\sum_i \sum_j{}' \hat{Y}_{ij}}{m}\right) - \left(\frac{\sum_i \sum_j{}^0 \hat{Y}_{ij}}{w}\right),$$

and the

unexplained portion of the mean wage disparity = $\Delta\bar{r}$ =

$$\left(\frac{\sum_i \sum_j{}' r_{ij}}{m}\right) - \left(\frac{\sum_i \sum_j{}^0 r_{ij}}{w}\right).$$

$\Delta\bar{r}$ represents that part of the mean wage disparity d left unexplained after stratified regression adjustment and is the proposed measure of discrimination in the general case. To assess the significance of $\Delta\bar{r}$, one considers exactly the same urn model as before for the random selection of residuals and computes the ratio

$$z = \frac{\Delta\bar{r}}{\left(\sum_{i=1}^{k} s_i^2 \frac{m_i w_i}{n_i}\right)^{1/2} \left(\frac{1}{m} + \frac{1}{w}\right)}, \qquad (2)$$

where now

$$s_i^2 = \begin{cases} \frac{1}{n_i - 1} \sum_{j=1}^{n} r_{ij}^2 & \text{for } n_i > 1 \\ 0 & \text{for } n_i \leqq 1 \end{cases} \qquad (i=1, \ldots, k).$$

If $(\Delta\bar{r})_i = (\Sigma' r_{ij}/m_i) - (\Sigma^0 r_{ij}/w_i)$ denotes the unexplained portion of the stratum specific mean wage disparity in stratum $i=1, \ldots, k$, then a little algebra provides the following computing formulas:

$$\Delta\bar{r} = \frac{\sum_{i=1}^{k} h_i(\Delta\bar{r})_i}{(mw/n)}$$

where $h_i = m_i w_i / n_i$, and (2) becomes

$$z = \frac{\sum_{i=1}^{k} h_i(\Delta\bar{r})_i}{(\sum_{i=1}^{k} h_i s_i^2)^{1/2}}.$$

Note that the unexplained portion of the mean wage disparity, $\Delta\bar{r}$, is generally not a weighted average of the stratum-specific $(\Delta\bar{r})_i$.

The role of regression in this approach is simply to provide an adjustment that can be used to represent fairly that part of the mean wage disparity accounted for by the differential occurrence of explanatory factors between men and women within each stratum.

III

HOW MANY REGRESSION SURFACES?

This section assumes a single stratum. With p explanatory variables $\underline{X} = (X^{(1)}, \ldots, X^{(p)})$, the usual sex-coefficient measure of shortfall is $\hat{\beta}$, the least squares estimate of β in the analysis of covariance model

$$Y = \alpha + \beta Z + \gamma_1 X^{(1)} + \ldots + \gamma_p X^{(p)} + e \tag{3}$$

where Z is the sex indicator (1 for men, 0 for women) and e is a random error term, assumed to satisfy

$$E(e \mid Z, X^{(1)}, \ldots, X^{(p)}) = 0$$

and

$$Var(e \mid Z, X^{(1)}, \ldots, X^{(p)}) = \sigma^2.$$

The quantity $\hat{\beta}$ is the distance between two estimated parallel hyperplanes representing the relationship between Y and $\underline{X}$ for men and women. A critical issue here is the justification of the parallelism assumption. A question not often addressed is: What does $\hat{\beta}$ measure when this assumption is false? Theorem 1 below and the discussion following it provide an interpretation of $\hat{\beta}$ even when the regression surfaces are not parallel. In such cases it is reasonable to fit two different regression surfaces, one for men and one for women, and to compare the corresponding regression coefficients on the X variables:

$$Y_1 = a_1 + c_1^{(1)}X^{(1)} + \ldots + c_1^{(p)}X^{(p)} + e_1 \quad \textit{for men}$$
$$Y_0 = a_0 + c_0^{(1)}X^{(1)} + \ldots + c_0^{(p)}X^{(p)} + e_0 \quad \textit{for women}.$$

A clear description of shortfall is more difficult to achieve in this case, for it must involve a careful examination of the relative positions of the two surfaces in different regions of a p-dimensional space. If one plane lies above the other for all relevant values of $\underline{X} = (X^{(1)}, \ldots, X^{(p)})$, an inference of significant disparity is fairly straightforward; however, when the planes cross for realizable values of the explanatory variables, any inference of disparity must be conditioned on the appropriate region. Elizabeth Scott has discussed plotting the residuals for men and women obtained by substituting the $\underline{X}$ values for men into the estimated women's equation, and vice versa.[9] A summary measure might then be one-half the difference between the men's mean residual and the women's mean residual, which might be called Scott's measure, although not explicitly proposed by Scott. Considered below is the relationship between $\Delta\bar{r}$, $\hat{\beta}$, and this measure.

A general criticism of both the sex-coefficient $\hat{\beta}$ and the comparison of the regression coefficients $\hat{c}_I^{(j)}$, $\hat{c}_0^{(j)}$ based on fitting a separate surface for men and women is that emphasis is placed on the *regression surfaces* rather than on the *actual persons* involved. Consider the hypothetical data below describing a group of three men and three women, their salaries, and the values of an explanatory variable X, seniority in years.

TABLE 1

individual	sex	seniority	salary ($)
1	F	2	10,000
2	F	4	10,000
3	F	6	10,000
4	M	6	10,000
5	M	7	15,000
6	M	8	20,000

Figure A shows the appropriate two-regression-line analysis in which it is clear that men receive $5,000 per year of seniority while women receive no such reward per year of seniority. Thus, on the basis of seniority regression coefficients, it appears that men and women are being treated differently. However, which individuals in the group can claim damages? Certainly, each person is doing at least as well on the basis of his or her own regression line as a person of similar seniority of the opposite sex. The average men's residual obtained from the women's line is (0 + 5,000 + 10,000)/3 = $5,000, whereas the average women's residual obtained from the men's line is (0 + 10,000 + 20,000)/3 = $10,000, indicating by Scott's measure that women are actually *ahead* of men. The analysis of covariance, indicated in figure B, shows just the opposite, namely a $2,000 sex-coefficient $\hat{\beta}$ favoring men. Thus, the *regression coefficients relate not to the actual employees, but to a hypothetical class of individuals who would be aggrieved only if they were to exist in the group of actual employees.* For completeness, observe that the method proposed in section II shows that of the mean wage disparity d = $5,000, all but $\Delta\bar{r}$ = $851.06 is explained by regression adjustment using a single line in which each additional year of seniority is rewarded with $1,382.98.[10] (Note that none of these measures would have much relevance to the discussion if it came to light that the employer's compensation scheme required six or more years of seniority before additional seniority was rewarded. The question would then shift to whether such a scheme has disparate impact because of past practices regarding hiring, firing, employee mobility and availability, and other aspects of employment.)

[9] E. Scott, Higher Education Salary Evaluation Kit 10-12 (1977) (available from the American Association of University Professors, Wash., D.C.).

[10] *See infra* Figure C, at 258.

FIGURE A. Separate regressions

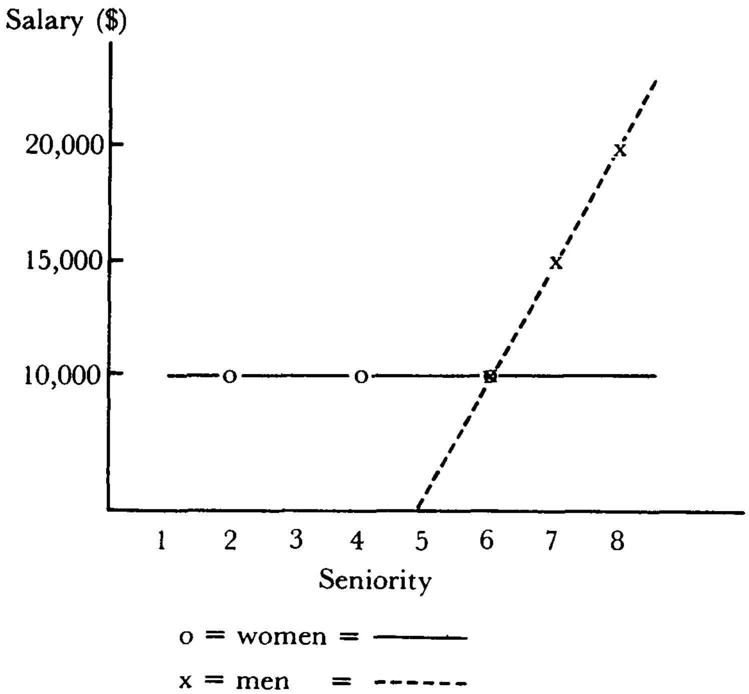


FIGURE B. Parallel regressions

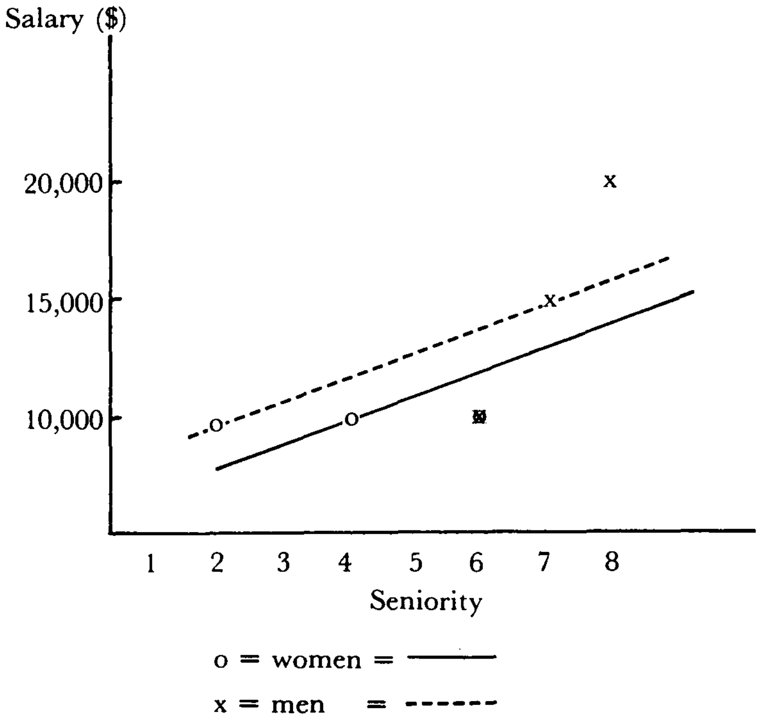
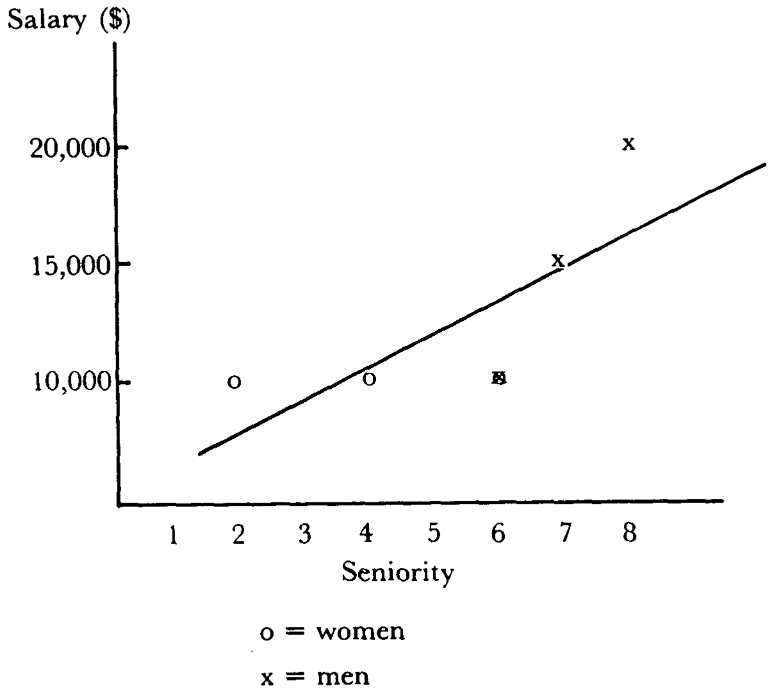


FIGURE C. One regression



The fact that $\Delta\bar{r}$ is substantially less than the sex-coefficient $\hat{\beta}$ in the example is not accidental, and is related to the fact that X = seniority is a good predictor of sex in this example. This is made explicit in the following

Theorem 1. Let $X^{(1)}, \ldots, X^{(p)}$ be explanatory variables, Z = sex (1 for men, 0 for women), and let $\hat{\beta}$ be the estimated sex-coefficient in model (3) using least squares. Assuming a single stratum, the unexplained portion of the mean wage disparity $\Delta\bar{r}$ is related to $\hat{\beta}$ by

$$\Delta\bar{r} = \hat{\beta}R^2{}_{Z\cdot\underline{X}}$$

where $R^2{}_{Z\cdot\underline{X}}$ is the coefficient of nondetermination of Z given $\underline{X} = (X^{(1)}, \ldots, X^{(p)})$, that is, one minus the squared multiple correlation coefficient of Z with $\underline{X}$.

Proof. Consider the following characterization of $\hat{\beta}$. First relate Y to $\underline{X}$ via

$$Y = a + c_1X^{(1)} + \ldots + c_pX^{(p)} + r$$

where $a, c_1, \ldots, c_p$ are least squares estimates and r denotes the residual used to compute $\Delta\bar{r}$, and relate Z to $\underline{X}$ via

$$Z = a' + c'_1X^{(1)} + \ldots + c'_pX^{(p)} + r'.$$

Then $\hat{\beta}$ is known to be the least squares estimate of β in the relation

$$r = \beta r' + e$$

(it is easily checked that e satisfies the normal equations, $0 = e\cdot 1 = e\cdot X^{(j)}$ for all j, and $0 = e\cdot r' = e\cdot(Z - (a' + \Sigma c'_jX^{(j)})) \Rightarrow e\cdot Z = 0$). The proof now follows from

$$\Delta\bar{r} = (\frac{1}{m}\Sigma' r_i) - (\frac{1}{w}\Sigma^0 r_i) = \frac{(\Sigma' r_i)}{mw/n} = \frac{\Sigma r_iZ_i}{\Sigma(Z_i - \bar{Z})^2}$$

$$= \frac{\Sigma r_i r'_i}{\Sigma(Z_i - \bar{Z})^2} \text{ since } \Sigma r_i = \Sigma r_i X_i^{(j)} = 0 \text{ for all } j$$

$$= \frac{(\Sigma r_i r'_i)\,(\Sigma r'^2_i)}{(\Sigma r'^2_i)\,\Sigma(Z_i - \bar{Z})^2} = \hat{\beta} R^2_{Z\cdot \underline{X}}\,.$$

One should observe that $\Delta\bar{r}$ diminishes the estimate $\hat{\beta}$ by the fraction of variation in Z not explained by the best linear predictor of Z based on $\underline{X}$. Heuristically, the more highly correlated sex is with the explanatory variables, the more $\Delta\bar{r}$ discounts $\hat{\beta}$. Put another way, the more separated men and women are on the basis of the explanatory variables, the less comparable they are on this basis and the smaller is the unexplained portion of the mean wage disparity based on the given adjustment formula for all persons.

The measure $\Delta\bar{r}$ may also be viewed in the following way. If sex is associated with salary after adjustment by $\underline{X}$, the residuals $r_1, \ldots, r_n$ will then be correlated with sex, and the magnitude of this correlation is of interest. In fact, it is easily seen that in the single stratum case, the ratio z in (2) is just $\sqrt{n-1}$ times the product-moment correlation of sex with the residuals.

This section concludes with a brief observation about the several measures of discrimination considered in the single stratum case. Let

$\mu_1 = \hat{\beta} =$ the sex-coefficient,

$\mu_2 = \Delta\bar{r} =$ the unexplained portion of the mean wage disparity,

$\mu_3 =$ Scott's measure,

and a fourth,

$\mu_4 =$ the average distance between the separate men's and women's regression surface at the observed $\underline{X}$ values, that is,

$$\mu_4 = \frac{1}{n}\sum_{i=1}^{n} (\hat{Y}_1(\underline{X}_i) - \hat{Y}_0(\underline{X}_i))$$

where

$\hat{Y}_1(\underline{X})$ is the estimated men's regression surface at a point $\underline{X}$

and

$\hat{Y}_0(\underline{X})$ is the estimated women's regression surface at a point $\underline{X}$.

For simplicity, consider the case of a single explanatory variable X. It is easily seen that each of these measures can be expressed as an adjusted mean wage disparity: for $i = 1, 2, 3, 4$ we have

$$\mu_i = (\bar{Y}_1 - \bar{Y}_0) - c_i(\bar{X}_1 - \bar{X}_0) \equiv \Delta\bar{Y} - c_i\,\Delta\bar{X}, \tag{4}$$

where

$c_1 = \hat{\gamma} =$ least squares estimate of the X coefficient in the analysis of covariance;

$c_2 = \hat{c} =$ least squares estimate of X without sex in the regression function;

$c_3 = \frac{1}{2}(\hat{c}_1 + \hat{c}_0) =$ the average of the separate regression on X coefficients;

and

$$c_4 = (\frac{w}{n})\hat{c}_1 + (\frac{m}{n})\hat{c}_0 \cdot$$

Since γ is a convex combination of $\hat{c}_1$ and $\hat{c}_0$ with weights given by

$$\frac{\underset{i}{\Sigma^1}(X_i - \bar{X}_1)^2}{\underset{i}{\Sigma^1}(X_i - \bar{X}_1)^2 + \underset{i}{\Sigma^0}(X_i - \bar{X}_0)^2} \text{ and } \frac{\underset{i}{\Sigma^0}(X_i - \bar{X}_0)^2}{\underset{i}{\Sigma^1}(X_i - \bar{X}_1)^2 + \underset{i}{\Sigma^0}(X_i - \bar{X}_0)^2},$$

the three measures μ_1, μ_3, μ_4 are similar in this respect. On the other hand, $\hat{c}$ is not a linear combination of $\hat{c}_1$ and $\hat{c}_0$ alone but a convex combination of $\hat{c}_1$, $\hat{c}_0$, and $\frac{\Delta\bar{Y}}{\Delta\bar{X}}$. In fact, using Theorem 1 and (4) one can see that

$$\hat{c} = \hat{\gamma} R^2_{Z \cdot X} + \frac{\Delta\bar{Y}}{\Delta\bar{X}} R^2_{Z(X)}.$$

In the usual case with $\Delta\bar{X} > 0$, $\hat{\beta} > 0$, Theorem 1 shows that $\hat{\gamma} < \hat{c}$, whence

$$\hat{\gamma} < \hat{c} < \frac{\Delta\bar{Y}}{\Delta\bar{X}} \cdot$$

Finally, the form of (4) indicates the effect of an arbitrary linear adjustment formula of the form

$$\hat{Y} = \bar{Y} + c(X - \bar{X})$$

on the measure $\Delta\bar{r}$.

IV

COMPARING SIGNIFICANCE TESTS WITH Z AND T

The following theorem gives the relationship between z and the usual t-statistic for the sex-coefficient.

Theorem 2. In a single stratum

$$z = t \frac{R_{Y \cdot ZX} R_{Z \cdot X}}{R_{Y \cdot X}} \left(\frac{n-1}{n-p-2}\right)^{1/2}$$

$$= \frac{t}{(1 + t^2/(n-p-2))^{1/2}} R_{Z \cdot X} \left(\frac{n-1}{n-p-2}\right)^{1/2}$$

where t is Student's t-statistic for testing the significance of $\hat{\beta}$ in model (3), and where

$$R^2_{Y \cdot \underline{X}} = 1 - R^2_{Y(\underline{X})} \text{ and } R^2_{Y \cdot ZX} = 1 - R^2_{Y(Z\underline{X})}$$

are obtained from the multiple correlation coefficients of Y with $\underline{X}$, and Y with (Z, $\underline{X}$), respectively.

Proof.

$$z = \Delta\bar{r}\{\Sigma(Z_i - \bar{Z})^2 / \Sigma r_i^2\}^{1/2} (n-1)^{1/2}$$

$$= \hat{\beta} R^2_{Z \cdot \underline{X}} \{\Sigma(Z_i - \bar{Z})^2 / \Sigma r_i^2\}^{1/2} (n-1)^{1/2} \text{ by Theorem 1.}$$

Now, the variance of $\hat{\beta}$ in the analysis of covariance model (3)[11] is

$$Var(\hat{\beta}) = \sigma^2/R^2_{Z\cdot\underline{X}}\Sigma(Z_i - \bar{Z})^2,$$

and when σ^2 is estimated by the unbiased estimate

$$\hat{\sigma}^2 = \Sigma\ (Y_i - \bar{Y})^2R^2_{Y\cdot Z\underline{X}}\ /\ (n-p-2)$$

one obtains the estimated

$$V\hat{a}r(\hat{\beta}) = \frac{\Sigma(Y_i - \bar{Y})^2\ R^2_{Y\cdot Z\underline{X}}\ /\ (n-p-2)}{\Sigma(Z_i - \bar{Z})^2\ R^2_{Z\cdot\underline{X}}}\ .$$

Thus, multiplying numerator and denominator of $\hat{\beta}$ by $V\hat{a}r^{1/2}(\hat{\beta})$ one obtains

$$z = t\ \frac{\{\Sigma(Y_i - \bar{Y})^2\ R^2_{Y\cdot ZX}\ /\ (n-p-2)\}^{1/2}}{\{\Sigma(Z_i - \bar{Z})^2\ R^2_{Z\cdot X}\}^{1/2}}\ R^2_{Z\cdot\underline{X}}\ \{\Sigma(Z_i - \bar{Z})^2\ /\ \Sigma r_i^2\}^{1/2}(n-1)^{1/2}\ .$$

$$= t\ \frac{R_{Y\cdot ZX}}{R_{Y\cdot X}}\ R_{Z\cdot\underline{X}}\ \left(\frac{n-1}{n-p-2}\right)^{1/2}.$$

The second equality of the theorem follows from the relation

$$t^2 = (R^2_{Y\cdot\underline{X}} - R^2_{Y\cdot Z\underline{X}})\ /\ (R^2_{Y\cdot Z\underline{X}}/(n-p-2)).$$

This completes the proof.

Since $R_{Y\cdot Z\underline{X}}/R_{Y\cdot\underline{X}} \leqq 1$, one sees that for large samples that $|z|<|t|$ by approximately the amount

$$R_{Y\cdot Z\underline{X}}R_{Z\cdot\underline{X}}\ /\ R_{Y\cdot\underline{X}}.$$

Now examine this from the viewpoint of the linear model (3). Under the null hypothesis H_0: $\beta = 0$, the factor

$$(R_{Y\cdot Z\underline{X}}/R_{Y\cdot\underline{X}}) \xrightarrow{P} 1 \text{ as } m \text{ and } w \longrightarrow \infty\ ,$$

so that when $R_{Z\cdot\underline{X}} = 1$, as in the no regression case, the two statistics are asymptotically equivalent. (The same is true, of course, if $R_{Z\cdot\underline{X}} \longrightarrow 1$.) However, if $R_{Z\cdot\underline{X}} \longrightarrow l < 1$, then the two statistics are not equivalent, even in the null case. The difference arises from the *nonexchangeability of the residuals* in the linear model, and is discussed further below.

In the several strata case, it is entirely possible for z in (2) to exceed the t-statistic for the sex-coefficient in an analysis of covariance model including stratum indicators as explanatory variables. This is true in part because the variance of the sum of residuals based on stratified random sampling may be very much smaller than the variance of the sum based on simple random sampling; thus, the value of z from (2) may substantially exceed the z that would be obtained from a single stratum analysis with stratum indicators as explanatory variables. If this latter value is not too much smaller than the corresponding t value, it is possible that the former could exceed the t-statistic. This fact underscores the benefit of stratification in the analysis.

Returning to the single stratum case, the result of theorem 2 may appear paradoxical in asserting that under H_0 for large n the z score, which is asymptotically

[11] *See* F. MOSTELLER & J. TUKEY, DATA ANALYSIS AND REGRESSION 342 (1977).

standard normal, is pointwise less than the t-statistic, which is also asymptotically standard normal. The explanation lies in the fact that in the linear model (3) under H_0, the residuals

$$\hat{e}_i = Y_i - (\hat{\alpha} + \hat{\gamma}_1 X_i^{(1)} + \ldots + \hat{\gamma}_p X_i^{(p)}) = r_i$$

are not exchangeable random variables, contradicting the assumption of equal likelihood for all subsets of residuals in the urn model. Although it is true that the residuals are uncorrelated with each $X^{(j)}$, there is still information about r_i in $(X_i^{(1)}, \ldots, X_i^{(p)})$. In fact, given model (3) with $\beta = 0$, the covariance matrix of $\underline{r}$ is

$$\sigma^2(I - X(X'X)^{-1}X'),$$

where X is the usual $n \times (p + 1)$ design matrix with columns $(\underline{1}, \underline{X}^{(1)}, \ldots, \underline{X}^{(p)})$, and I is the $n \times n$ identity matrix. In the case $p = 1$, for example, the variance of the i^{th} residual r_i is

$$Var(r_i) = \sigma^2(1 - \frac{1}{n} - \frac{(X_i - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}),$$

indicating that residuals corresponding to extreme X values are stochastically smaller in magnitude than those corresponding to central values of X; that is, *the least squares line fits the extreme points better than it does the middle points.*

In the urn model the residuals are treated as observed data relative to the given adjustment formula. One can view the differences between the usual linear model and the urn model as arising from changing one's view of the residuals from that of correlated estimates of unobservable independent errors in the linear model to that of primary data in the finite urn population, with exchangeability instead of correlation. To put this another way, if in the linear model the independent errors e were actually observable and known, that is, if the assumed true regression equation were known, the analysis would then be exact, for in this case the exchangeability of the residuals would follow from the independence of the errors. Since this article views the regression adjustment with least squares coefficients as *known* in the given finite population, the urn model approach is appealing.

As to which method is "correct," the answer depends on the assumptions one is willing to make. If the linear model (3) is assumed to hold, then the urn model in the regression case is technically incorrect. On the other hand, the nonexchangeability of the residuals makes sense only with respect to some notion of replicability of the data with random errors, as in model (3), which must be considered suspect in nonexperimental settings. The urn model does not assume the random error concept and is therefore universally valid as a standard of comparison. Of course, the urn model makes use of the notions of replicability and random variation, but the site of these concepts has been shifted away from the error term to an urn experiment where replicability and random variation have a concrete and easily understood meaning.

This being said, one may nevertheless inquire to what extent the urn model may be modified to bring it more into line with the usual linear model. One method for doing this has been discussed by David Freedman and David Lane[12] and independently by Albert Beaton.[13]. They consider the permutation distribution of the estimated sex-coefficient $\hat{\beta}$ obtained from 'mock' data values $\tilde{Y}_i = \hat{Y}_i + r_{\pi(i)}$ where π is a permutation of $(1, \ldots, n)$ and where $\hat{Y}_i$ is the predicted value of Y_i based on $\underline{X}$. They show that the proportion of permutations for which the t-statistic for $\hat{\beta}$ (obtained from the mock data) exceeds the actual observed value t_{obs} is approximately the nominal value $P(t < t_{obs})$.[14] Although it reaffirms the usual methodology and interprets it in a nonstochastic setting, Freedman and Lane's model is less appealing than the urn model because it is hard to see the relevance of the distribution of $\hat{\beta}$ from mock values of the data. In addition, the residuals are not equally likely to attach themselves to values $\hat{Y}_i$ with any permutation in the usual model, for the reasons stated above.

The use of stratification in this analysis can be seen as a device to mitigate the nonexchangeability of the residuals in the classical model with a single stratum. In the no regression case there are exchangeable residuals within strata, and the urn model is consistent with the classical model. This suggests a possible approach to the regression case in which the domain of the explanatory variables is subdivided into smaller regions in each of which the assumption of exchangeability is approximately correct. This subdivision simply amounts to a further stratification of the data before regression adjustment. Alternatively, the adjustment formula may be based entirely on *discretized* explanatory variables for which the assumption of within-stratum exchangeability is consistent with the corresponding linear model. A different approach, not pursued here, would be to use a method of adjustment other than least squares that renders the residuals exchangeable, presumably by deemphasizing the role of the extreme points.

V

DISCUSSION

This article's approach to testing the significance of group salary differences is based on a randomization test applied to the residuals from a specific regression adjustment formula. This method provides an interpretable measure of disparity whose properties have been discussed in the preceding sections. This section explains why, in the context of employment discrimination litigation, randomization tests may be preferable to tests based on classical sampling theory.

The standard multiple linear regression model assumes the existence of a population of values from which the data at hand were drawn by a random sampling process. This assumption is appropriate only in certain circumstances. For example, the relevant population may consist of an actual large employee population, consisting of all N people satisfying given criteria during a specified time period. The values of observable quantities are considered fixed for each

[12] D. Freedman & D. Lane, Significance Testing in a Nonstochastic Setting (1979) (Univ. of Calif.-Berkeley Statistics Dep't Tech. Rep. #317).

[13] Beaton, *Salvaging Experiments: Interpreting Least Squares in Non-Random Samples,* in COMPUTER SCIENCE & STATISTICS: TENTH ANNUAL SYMPOSIUM ON THE INTERFACE 137 (1978).

[14] D. Freedman & D. Lane, *supra* note 12, at 10-13; Beaton, *supra* note 13, at 140-42.

employee, and the variation in these values from person to person constitutes "population variability." Random variation occurs only because of the assumption that a sample of n employees actually under consideration forms a random sample chosen without replacement from the larger population of size N. This means that one should be willing to treat all subsets of n employees in the population as equally likely to have comprised the actual observed data set. This assumption is appropriate when actual random sampling was used to generate the data.

Random sampling plays a central role in statistical analysis not because it guarantees that the sample obtained will be "representative" of the population, as is sometimes thought, but rather because it establishes a well-defined probability space, with equally likely outcomes, thereby making statistical inferences possible. In fact, the first step in any statistical analysis is to embed the observed data in an appropriate probability space. How to specify this space properly, so that it accurately and realistically reflects the class of outcomes that might have been observed, is the central problem considered here.

If the population variability is such that, within any class having fixed values of the explanatory factors, the dependent variable has a mean value that is a linear function of the explanatory variables, then the second assumption of the standard linear model is satisfied. The third assumption, that the sampled values are statistically independent, will be approximately true if the fraction of the population sampled is small or if sampling is actually done with replacement. Two other assumptions are often made for convenience, namely that the population variability within any class having fixed explanatory factors is constant and is approximately normal in distribution. These last two assumptions can be checked by an examination of the residuals, and unless they are checked there is a threat to the validity of the conclusions drawn.

Returning to the notion of the relevant population, we observe that the relevant population may not always be an actual set of people but may be a hypothetical set such as all "potential" employees or all employees "past, present, and future." Without a definite sampling frame, it becomes very difficult if not impossible to justify the random sampling assumption. It is clear that in real situations a given group of employees is a very specially selected, nonrandom collection of individuals. This point is crucial, for example, in appreciating the importance of applicant flow data and the potential irrelevance of census data when discrimination in hiring is at issue.

At the other extreme, the observed set of employees may actually *exhaust* the relevant population; in this case the probability space degenerates to a single outcome, and there are no longer any nontrivial statistical inferences to be made. This leads to the often-quoted statement that in a completely observed finite population any differences between groups are significant. While true in a strict sampling sense, this narrow attitude is not useful because it fails to allow for any variation in the way things might have been. To move a step closer to this goal one must either widen one's definition of the relevant population or else revise one's notion of the probability space. These options are explored below. Note that when one uses a population sampling model, the result of the statistical analysis is a set of (hopefully correct) statements not merely about the sample but about the entire population, even though the legal relevance of the larger set may be questioned.

With a rather abrupt change of viewpoint, the population is often conceived to be an infinite collection of "errors," such that each employee's salary is regarded as a deterministic value specified by the linear model, plus a random error term. This approach is appropriate in physical applications where random measurement error perturbs the observation of the "true" or mean value. In the employment context, however, it is difficult to accept the idea that an employer determines salary by adding a random error to a deterministic value that is calculated from a specified linear model. Such a model is unrealistic, and even if an employer did operate in this fashion, budgetary constraints would in most cases render the error distributions statistically dependent and nonnormal. To circumvent this objection, the view of the error term is revised so that it now represents the net effect of many deterministic but inaccessible factors beyond those specified in the model. This view is perhaps the most valid interpretation of the population sampling-of-errors approach, and it does accomplish an embedding of the observed results in a probability space that may come close to representing a set of relevant possible outcomes. However, much vagueness concerning the error population remains, since it is a largely hypothetical and speculative construction. Nor is it clear how one could ever verify or test the assumption of random sampling. In real employment situations one often finds that the error distributions differ on a case-by-case basis that reflects the different factors that are relevant to different employees. And, as always, unless the net effect of the unobservable factors is uncorrelated with the specified variables in the linear model, there will be bias in the estimation of the model parameters.

Generally, population sampling models have been successfully applied in the natural sciences because in an experimental setting the probability space becomes known empirically through replication. By replication one may learn the true extent of population variability and the structure of error distributions. In fact, the frequentist interpretation of probability is stated in terms of a sequence of replications of a sampling experiment. In observational studies of unique events and outcomes, however, one does not have the opportunity to verify these assumptions about the probability space, so that statistical arguments based on population sampling models are often not compelling. To some extent a specialist may draw on his expertise in a given area to supply the missing empirical validation of his assumptions. Thus, a physicist may know the nature of his measurement error distributions from theoretical considerations without extensive checking, although checking should be done in any novel situation. Economists and labor market experts often seem willing to assume a standard regression model without checking, possibly because they feel that they "know" the probability space through their experience in other similar cases. It should be made clear, though, that the expert's analysis ultimately rests on his opinion that the particular case at hand is a random realization of a more general phenomenon. It is then for the court to decide on the relevance of the larger process to the particular case at issue.

Regression models have been used convincingly in the natural sciences for another reason—in an experimental setting the researcher can actually manipulate the explanatory factors and control confounding factors. This manipulation is not possible in observational studies. Furthermore, those factors beyond the researcher's direct control can be prevented from causing serious bias by randomizing experimental units across treatments or conditions. These standard devices are employed in good experimental design to assure the validity of statements that refer to the partial effects of some variables while controlling for others. This type of control is not possible in the context of employment studies, and since one cannot "control" the effect of factors such as sex and education on salary, one needs to be wary of broad statements that such confounding factors have been successfully controlled. What one really means to assert is that an effort at adjustment has been made in these comparisons to account for uncontrolled differences between groups in one or more important factors. Thus, an emphasis is placed on the use of regression to provide adjustment formulas rather than estimates of partial effects.

All of these problems hamper the straightforward application of standard linear regression models to the analysis of employment data in litigation. This is not to say that the standard model can never be valid, but that defending the validity of inferences is considerably more difficult with this approach than with the urn model approach.

In the urn model approach the probability space is the set of all possible permutations of sex with the unexplained portion of salary. The only assumption necessary under the null hypothesis that sex has no direct influence on salary is that each assignment of sexes to the salary residuals is equally likely. The analysis in entirely conditional on the observed salary data and the associated adjustment formula; the set of possible outcomes in which one embeds the observed data involves neither hypothetical salaries nor error distributions of any kind. Moreover, in the urn model approach there is no estimate of population parameters, nor is there a model of the detailed way in which the employer has set salaries (although introducing one may be desirable to produce a realistic adjustment formula). This approach does not even assume, for example, that the employer adjusts salaries on the basis of explanatory factors with the same adjustment formula that is used in court. Put simply, the concern here is not on *what* salaries an employer has set for his employees but on *who* gets the large, unexplained portions of salary and whether or not there is any sex bias in the assignment. This is a sensible approach to take for the specific goal of testing for a "pattern and practice of discrimination."

VI

CONCLUSION

This article has proposed a method for statistical studies of wage disparities using regression adjustment methods that depend on a minimal number of assumptions for their validity. The urn models are seen to provide reasonable standards of comparison for the statistical assessment of unexplained wage disparities. Since the models need not give results equivalent to those of the customary linear random error models, they force one to evaluate the role of causal modeling in the context of employment. Given the general paucity of causal models and the difficulties associated with them, it is recommended that serious consideration be given to the urn model approach in the absence of any more penetrating and robust analysis.