Sex Bias in Graduate Admissions: Data from Berkeley

Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation.

P. J. Bickel, E. A. Hammel, J. W. O'Connell

Determining whether discrimination because of sex or ethnic identity is being practiced against persons seeking passage from one social status or locus to another is an important problem in our society today. It is legally important and morally important. It is also often quite difficult. This article is an exploration of some of the issues of measurement and assessment involved in one example of the general problem, by means of which we hope to shed some light on the difficulties. We will proceed in a straightforward and indeed naive way, even though we know how misleading an unsophisticated approach to the problem is. We do this because we think it quite likely that other persons interested in questions of bias might proceed in just the same way, and careful exposure of the mistakes in our discovery procedure may be instructive.

Data and Assumptions

The particular body of data chosen for examination here consists of applications for admission to graduate study at the University of California, Berkeley, for the fall 1973 quarter. In the admissions cycle for that quarter, the Graduate Division at Berkeley received approximately 15,000 applications, some of which were later withdrawn or transferred to a different proposed entry quarter by the applicants. Of the applications finally remaining for the fall 1973 cycle 12,763 were sufficiently complete to permit a decision to admit or to deny admission. The question we wish to pursue is whether the decision to admit or to deny was influenced by the sex of the applicant. We cannot know with any certainty the influences on the evaluators in the Graduate Admissions Office, or on the faculty reviewing committees, or on any other administrative personnel participating in the chain of actions that led to a decision on an individual application. We can, however, say that if the admissions decision and the sex of the applicant are statistically associated in the results of a series of applications, we may judge that *bias* existed, and we may then seek to find whether *discrimination* existed. By "bias" we mean here a pattern of association between a particular decision and a particular sex of applicant, of sufficient strength to make us confident that it is unlikely to be the result of chance alone. By "discrimination" we mean the exercise of decision influenced by the sex of the applicant when that is immaterial to the qualifications for entry.

The simplest approach (which we shall call approach A) is to examine the aggregate data for the campus. This approach would surely be taken by many persons interested in whether bias in admissions exists on any campus. Table 1 gives the data for all 12,763 applications to the 101 graduate departments and interdepartmental graduate majors to which application was made for fall 1973 (we shall refer to them all as departments). There were 8442 male applicants and 4321 female applicants. About 44 percent of the males and about 35 percent of the females were admitted. Just this kind of simple calculation of proportions impels us to examine the data further. We will pursue the question by using a familiar statistic, chi-square. As already noted, we are aware of the pitfalls ahead in this naive approach, but we intend to stumble into every one of them for didactic reasons.

We must first make clear two assumptions that underlie consideration of the data in this contingency table approach. Assumption 1 is that in any given discipline male and female applicants do not differ in respect of their intelligence, skill, qualifications, promise, or other attribute deemed legitimately pertinent to their acceptance as students. It is precisely this assumption that makes the study of "sex bias" meaningful, for if we did not hold it any differences in acceptance of applicants by sex could be attributed to differences in their qualifications, promise as scholars, and so on. Theoretically one could test the assumption, for example, by examining presumably unbiased estimators of academic qualification such as Graduate Record Examination scores, undergraduate grade point averages, and so on. There are, however, enormous practical difficulties in this. We therefore predicate our discussion on the validity of assumption 1.

Assumption 2 is that the sex ratios of applicants to the various fields of graduate study are not importantly associated with any other factors in admission. We shall have reason to challenge this assumption later, but it is crucial in the first step of our exploration, which is the investigation of bias in the aggregate data.

Tests of Aggregate Data

We pursue this investigation by computing the expected frequencies of male and female applicants admitted and denied, from the marginal totals of Table 1, on the assumption that men and women applicants have equal chances of admission to the university (that is, on the basis of assumptions 1 and 2). This computation, also given in Table 1, shows that 277 fewer women and 277 more men were admitted than we would have expected under the assumptions noted. That is a large number, and it is unlikely that so large a bias to the disadvantage of women would occur by chance alone. The chi-square value for this table is 110.8, and the probability of a chi-square that large (or larger) under the assumptions noted is vanishingly small.

We should on this evidence judge

Dr. Bickel is professor of statistics, Dr. Hammel is professor of anthropology and associate dean of the Graduate Division, and Mr. O'Connell is a member of the data processing staff of the Graduate Division, at the University of California, Berkeley 94720.

that bias existed in the fall 1973 admissions. On that account, we should look for the responsible parties to see whether they give evidence of discrimination. Now, the outcome of an application for admission to graduate study is determined mainly by the faculty of the department to which the prospective student applies. Let us then examine each of the departments for indications of bias. Among the 101 departments we find 16 that either had no women applicants or denied admission to no applicants of either sex. Our computations, therefore, except where otherwise noted, will be based on the remaining 85. For a start let us identify those of the 85 with bias sufficiently large to occur by chance less than five times in a hundred. There prove to be four such departments. The deficit in the number of women admitted to these four (under the assumptions for calculating expected frequencies as given above) is 26. Looking further, we find six departments biased in the opposite direction, at the same probability levels; these account for a deficit of 64 men.

These results are confusing. After all, if the campus had a shortfall of 277 women in graduate admissions, and we look to see who is responsible, we ought to find somebody. So large a deficit ought not simply to disappear. There is even a suggestion of a surplus of women. Our method of examination must be faulty.

Some Underlying Dependencies

We have stumbled onto a paradox, sometimes referred to as Simpson's in this context (*1*) or "spurious correlation" in others (*2*). It is rooted in the falsity of assumption 2 above. We have assumed that if there is bias in the proportion of women applicants admitted it will be because of a link between sex of applicant and decision to admit. We have given much less attention to a prior linkage, that between sex of applicant and department to which admission is sought. The tendency of men and women to seek entry to different departments is marked. For example, in our data almost two-thirds of the applicants to English but only 2 percent of the applicants to mechanical engineering are women. If we cast the application data into a 2×101 contingency table, distinguishing department and sex of applicants, we find this table has a chi-square of 3091 and that the probability of obtaining a chi-square value that large or larger by chance is about zero. For the 2×85 table on the departments used in most of the analysis, chi-square is 3027 and the probability about zero. Thus the sex distribution of applicants is anything but random among the departments. In examining the data in the aggregate as we did in our initial approach, we pooled data from these very different, independent decision-making units. Of course, such pooling would not nullify assumption 2 if the different departments were equally difficult to enter. We will address ourselves to that question in a moment.

Table 1. Decisions on applications to Graduate Division for fall 1973, by sex of applicant—naive aggregation. Expected frequencies are calculated from the marginal totals of the observed frequencies under the assumptions (1 and 2) given in the text. $N = 12{,}763$, $\chi^2 = 110.8$, d.f. $= 1$, $P = 0$ (*18*).

Applicants	Outcome				Difference	
	Observed		Expected			
	Admit	Deny	Admit	Deny	Admit	Deny
Men	3738	4704	3460.7	4981.3	277.3	− 277.3
Women	1494	2827	1771.3	2549.7	− 277.3	277.3

Let us first examine an alternative to aggregating the data across the 85 departments and then computing a statistic—namely, computing a statistic on each department first and aggregating those. Fisher gives a method for aggregating the results of such independent experiments (*3*). If we apply his method to the chi-square statistics of the 85 individual contingency tables, we obtain a value that has a probability of occurrence by chance alone, that is, if sex and admission are unlinked for any major, of about 29 times in 1000 (*4*). Another common aggregation procedure, proposed to us in this context by E. Scott, yields a result having a probability of 6 times in 10,000 (*5*). This is consistent with the evidence of bias in some direction purportedly shown by Table 1. However, when we examine the *direction* of bias, the picture changes. For instance, if we apply Fisher's method to the *one-sided* statistics, testing the hypothesis of no bias or of bias in favor of women, we find that we could have obtained a value as large as or larger than the one observed, by chance alone, about 85 times in 100 (*6*).

Our first, naive approach of examining the aggregate data, computing expected frequencies under certain assumptions, computing a statistic, and deciding therefrom that bias existed in favor of men has now been cast into doubt on at least two grounds. First, we could not find many biased decision-making units by examining them individually. Second, when we take account of the differences among departments in the proportions of men and women applying to them and avoid this problem by computing a statistic on each department separately, and aggregating those statistics, the evidence for campus-wide bias in favor of men is extremely weak; on the contrary, there is evidence of bias in favor of women.

The missing piece of the puzzle is yet another fact: not all departments are equally easy to enter. If we cast the data into a 2×101 table, distinguishing department and decision to admit or deny, we find that this table has a chi-square value of 2195, with an associated probability of occurrence by chance (under assumptions 1 and 2) of about zero, showing that the odds of gaining admission to different departments are widely divergent. (For the 2×85 table chi-square is 2121 and the probability about zero.) Now, these odds of getting into a graduate program are in fact strongly associated with the tendency of men and women to apply to different departments in different degree. *The proportion of women applicants tends to be high in departments that are hard to get into and low in those that are easy to get into.* Moreover this phenomenon is more pronounced in departments with large numbers of applicants. Figure 1 is a scattergram of proportion of applicants that are women plotted against proportion of applicants that are admitted. The association is obvious on inspection although the relationship is certainly not linear (*7*). If we use a weighted correlation (*8*) as a measure of the relationship for all 85 departments in the plot we obtain $\hat{\rho} = .56$. If we apply the same measure to the 17 departments with the largest numbers of applicants (accounting for two-

thirds of the total population of applicants) we obtain $\hat{p} = .65$, while the remaining 68 departments have a corresponding $\hat{p} = .39$. The significance of $\hat{p}$ under the hypothesis of no association can be calculated. All three values obtained are highly significant.

The effect may be clarified by means of an analogy. Picture a fishnet with two different mesh sizes. A school of fish, all of identical size (assumption 1), swim toward the net and seek to pass. The female fish all try to get through the small mesh, while the male fish all try to get through the large mesh. On the other side of the net all the fish are male. Assumption 2 said that the sex of the fish had no relation to the size of the mesh they tried to get through. It is false. To take another example that illustrates the danger of incautious pooling of data, consider two departments of a hypothetical university—machismatics and social warfare. To machismatics there apply 400 men and 200 women; these are admitted in exactly equal proportions, 200 men and 100 women. To social warfare there apply 150 men and 450 women; these are admitted in exactly equal proportions, 50 men and 150 women. Machismatics admitted half the applicants of each sex, social warfare admitted a third of the applicants of each sex. But about 73 percent of the men applied to machismatics and 27 percent to social warfare, while about 69 percent of the women applied to social warfare and 31 percent to machismatics. When these two departments are pooled and expected frequencies are computed in the usual way (with assumption 2), there is a deficit of about 21 women (Table 2). A discrepancy in that direction that large or larger would be expectable less than 2 percent of the time by chance; yet both departments were seen to have been absolutely fair in dealing with their applicants.

Table 2. Admissions data by sex of applicant for two hypothetical departments. For total, $\chi^2 = 5.71$, d.f. $= 1$, $P = 0.19$ (one-tailed).

Applicants	Outcome				Difference	
	Observed		Expected			
	Admit	Deny	Admit	Deny	Admit	Deny
			Department of machismatics			
Men	200	200	200	200	0	0
Women	100	100	100	100	0	0
			Department of social warfare			
Men	50	100	50	100	0	0
Women	150	300	150	300	0	0
			Totals			
Men	250	300	229.2	320.8	20.8	− 20.8
Women	250	400	270.8	379.2	− 20.8	20.8

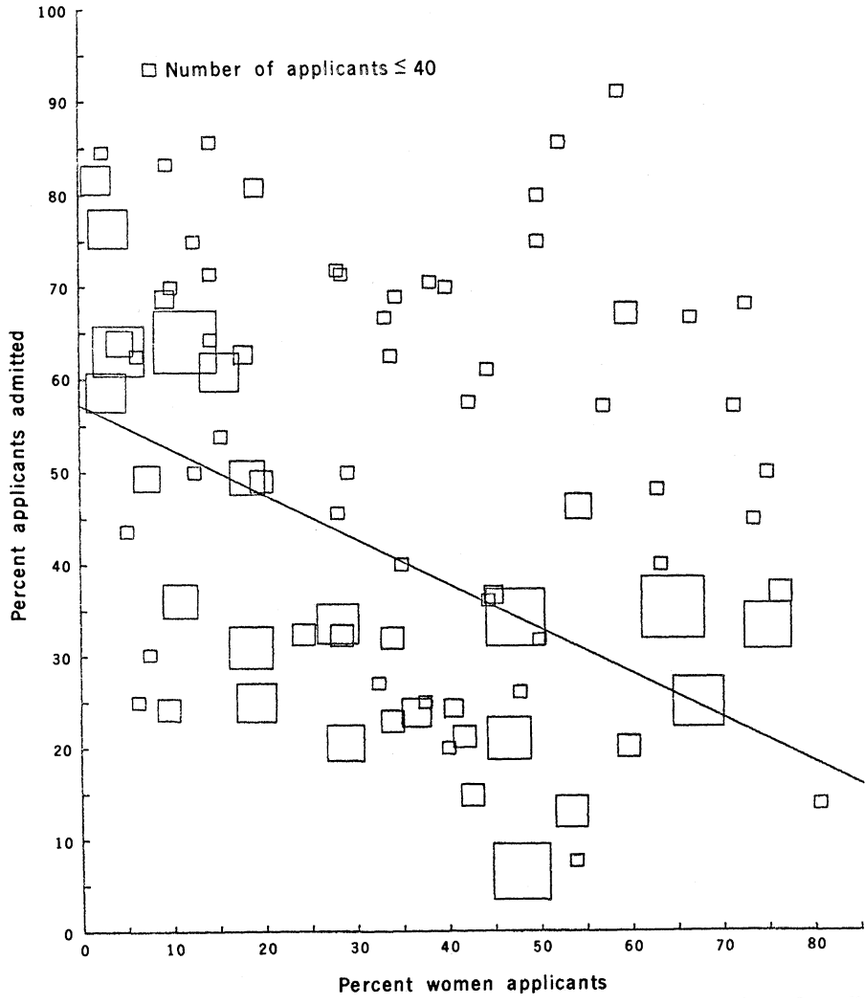

Fig. 1. Proportion of applicants that are women plotted against proportion of applicants admitted, in 85 departments. Size of box indicates relative number of applicants to the department.

The creation of bias in our original situation is, of course, much more complex, since we are aggregating many tables. It results from an interaction of the three factors, choice of department, sex, and admission status, whose broad outlines are suggested by our plot but which cannot be described in any simple way.

In any case, aggregation in a simple and straightforward way (approach A) is misleading. More sophisticated methods of aggregation that do not rely on assumption 2 are legitimate but have their difficulties. We shall have more to say on this later.

Disaggregation

The most radical alternative to approach A is to consider the individual graduate departments, one by one. However, this approach (which we may call approach B) also poses difficulties. Either we must sample randomly from the different departments, or we must take account of the probability of obtaining unusual sex ratios of admittees by chance in a number of simultaneously conducted independent experiments. That is, in examining 85 separate departments at the same time for evidence of bias we are conducting 85 simultaneous experiments,

and in that many experiments the probability of finding some marked departures from expected frequencies "just by chance" is not insubstantial. The department with the strongest bias against admitting women in the fall 1973 cycle had a bias of sufficient magnitude to be expectable by chance alone only 69 times in 100,000. If we had selected that department for examination on a random basis, we would have been convinced that it was biased. But we did not so select it; we looked at 85 departments at once. The probability of finding a department that biased against women (or more biased) by chance alone in 85 simultaneous trials is about 57 times in 1000. Thus that particular department is not quite so certainly biased as we might have first believed, .057 being a very much larger number than .00069, although still a small enough probability to warrant a closer look. This department was the worst one in respect of bias against women in admissions; the probability of finding departments less biased by chance alone is of course greater than .057. We can also examine events in the other direction. The department most biased against men had a bias sufficiently large to be expectable by chance alone about 20 times in a million, and the chance of finding a department that biased (or more biased) in that direction by chance alone in 85 simultaneous trials (*9*) is about .002.

There is a further difficulty in approach B. Although it makes a great deal of sense to examine the individual departments that are in fact the independent decision-making entities in the graduate admissions process, some of them are quite small, and even in some that are of ordinary size the number of women applying is very small. Calculation of the probability of observed deviations from expected frequencies can be carried out for such units, but when the numbers involved are very small the evidence for deciding whether there is no bias or gross bias is really worthless (*10*). This defect is evident not only in approach B but also if we use some reasonable method of aggregation of test statistics to avoid the pitfalls of approach A such as that of Fisher, or even the approach we suggest below. That is, large biases in small departments or in departments with small numbers of women applicants will not influence a reasonable aggregate measure appreciably.

Pooling

The difficulty we face is not only technical and statistical but also administrative. In some sense the campus is a unit. It operates under general regulations concerning eligibility for admission and procedures for admission. It is a social community that shares certain values and is subject to certain general influences and pressures. It is identifiable as a bureaucratic unit by its own members and also by external agencies and groups. It is, as a social and cultural unit, accountable to its various publics. For all these reasons it makes sense to ask the question, Is there a campus bias by sex in graduate admissions? But this question raises serious conceptual difficulties. Is campus bias to be measured by the net bias across all its constituent subunits? How does one define such a bias? For any definition, it is easy to imagine a situation in which some departments are biased in one direction and other departments in another, so that the net bias of the campus may be zero even though very strong biases are apparent in the subunits. Does one look instead at the outliers, those departments that have divergences so extreme as to call their particular practices into question? How extreme is extreme in such a procedure, and what does one do about units so small as to make such assessment meaningless?

We believe that there are no easy answers to these questions, but we are prepared to offer some suggestions. We propose that examination of campus bias must rest on a method of estimation of expected frequencies that takes into account the falsity of assumption 2 and the apparent propensity of women to apply to departments that are more difficult to enter.

We reanalyze Table 1, using all the data leading to it, by computing the expected frequencies differently than in approach A, since we now know the assumptions underlying that earlier computation to be false. We estimate the number of women expected to be admitted to a department by multiplying the estimated probability of admission of any applicant (regardless of sex) to that department by the number of women applying to it. Thus, if the chances of getting into a department were one-half for all applicants to it, and 100 women applied, we would expect 50 women to be admitted if they were being treated just like the men. We do this computation for each department separately, since each is likely to have a different probability of admission and a different number of women applying, and we sum the results to obtain the number of women expected to be admitted for the campus as a whole (*11*). This estimate proves to be smaller by 60 than the number of women observed to have been admitted (Table 3).

The computation of Table 3 is as follows: For a four-cell contingency table of the following format:

	Admit	Deny
Men	a_i	b_i
Women	c_i	d_i

the particular cell of interest is c_i, containing the number of women admitted. The expected frequency under the hypothesis of no bias is $E = w_i p_i = (c_i + d_i)(a_i + c_i)/N_i$, where N_i is the total of applicants to department i. The observed number, O, is the number in c_i. The difference between these two quantities, $O - E$, summed over n departments is

$$\sum_{i=1}^{n}(O - E) = DIFF$$

Then,

$$\chi^2 = \frac{(DIFF)^2}{\sum_{i=1}^{n}(a_i + b_i)(a_i + c_i)(c_i + d_i)(b_i + d_i)/N_i^2(N_i - 1)}$$

with d.f. = 1. Ninety-six departments were included in the computation, since 5 of the total 101 each had only 1 applicant. If $N_i - 1$ is replaced by N in the denominator, all 101 departments can be included, yielding $\chi^2 = 8.61$; $O - E$ remains 60.1 and the expected and observed female admittees are each increased by 1. (This statistic makes it possible to include contingency tables having an empty cell, so that no information is lost; there is thus an advantage over methods that pool the chi-square values from a set of contingency tables.)

The probability that an observed bias this large or larger in *favor* of women might occur by chance alone (under these new assumptions) is .0016; the probability of its occurring if there were actual discrimination against women is, of course, even smaller. This is consistent with what we found using Fisher's approach and aggregating the test statistics: there is evidence of bias in favor of women. [The test used here was proposed in

another context by Cochran (*12*) and Mantel and Haenszel (*13*).]

We would be remiss if we did not point out yet another pitfall of approach A. Whereas the highly significant values of the Mantel-Haenszel or Fisher statistics just mentioned for 1973 are evidence that there is bias in favor of women, the low values obtained in other years (see below) do not indicate that every department was operating more or less without bias. Such low values could equally well arise as a consequence of cancellation. We illustrate with the hypothetical departments of machismatics and social warfare. If machismatics admitted 250 men and 50 women, creating a shortfall of 50 women, while social warfare admitted 200 women and no men, creating an excess of 50 women, the aggregate measure of bias we have introduced would be zero. We only argue that if an aggregate measure of bias is wanted the one we propose is reasonable. Of course, if we combine two-sided statistics by the Fisher method this phenomenon does not occur.

We would conclude from this examination that the campus as a whole did not engage in discrimination against women applicants. This conclusion is strengthened by similarly examining the data for the entire campus for the years 1969 through 1973. In 1969 the number of women admitted exceeded the expected frequency by 24; the probability of a deviation of this size or larger in *either* direction by chance alone is .196. In 1970 there were four fewer women admitted than expected, the probability of chance occurrence being .833. In 1971 there were 25 more women than expected, with a probability of .249. In 1972 there were seven more women than expected, the probability being .709. For 1973 as shown above the deviation was an excess of 60 women over the expected number: the probability of a chance deviation that large or larger in either direction is .003. These data suggest that there is little evidence of bias of any kind until 1973, when it would seem significant evidence of bias appears, in favor of women. This conclusion is supported by all the other measures we have examined. For instance, pooling the chi-square statistics by Fisher's method yields a probability of .99 in 1969, 1970, and 1971, a probability of .55 in 1972, and a probability of .029 in 1973 (*14*).

We may also take approach B and look for individual department outliers. Because the numbers of women students applying to some of them in any one year are often small, we aggregated the data for each department over the 5-year span, using the method just explained. (This procedure of course hides the kind of change that the aggregating approach reveals when pursued through time, but it enables us to focus on possible "offenders" in either direction in a campus that is on the average behaving itself.) During the 5-year period there were 94 units that had at least one applicant of each sex and admitted at least one applicant and denied admission to at least one in at least one year. Two of the 94 units, one in the humanities and one in the professions, show a divergence from chance expectations sufficient to arouse interest. One of these admitted 16 fewer women than expected over 5 years, a shortfall of 29 percent; the probability of such a result by chance alone in 94 trials is about .004. The other unit admitted 40 fewer women than expected over the 5-year period, a shortfall of 7 percent, with a probability in 94 trials of about .019. The next most likely result by chance was at a level of .094 and the next after that at .188. Conversely there were two units significantly biased in the opposite direction, with chance probabilities of occurrence of .033 and .047, accounting for a combined shortfall of 50 men, 13 and 24 percent respectively of the expected frequencies in the individual units.

Table 3. Sum of expected departmental outcomes of women's applications compared with sum of observed outcomes, Graduate Division, Berkeley, fall 1973. $\chi^2 = 8.55$, d.f. $= 1$, $P = .003$ (two-tailed).

Expected female admittees	1432.9
Observed female admittees	1493.0
Difference $(O - E)$	60.1

The kinds of statistics we may wish to use in examination of individual departments may differ from those employed in these general screening processes. For example, in one of the cases of a shortfall of women cited above, it seems likely that an intensified drive to recruit minority group members caused a temporary drop in the proportion of women admitted, since most of the minority group admittees were males. In most of the cases involving favored status for women it appears that the admissions committees were seeking to overcome long-established shortages of women in their fields. Overall, however, it seems that the admissions procedure has been quite evenhanded. Where there are divergences from the expected frequencies they are usually small in magnitude (although they may constitute a substantial proportion of the expected frequency), and they more frequently favor women than discriminate against them.

More General Issues

We have already explained why assumption 1—the equivalence of academic qualifications of men and women applicants—is necessary to the statistical examination of bias in admissions. But the assumption is clearly false in its most extensive sense; there are areas of graduate study that men and women simply have not hitherto been equally prepared to enter. One of the principal differentiators is preparation in mathematics, which is prerequisite in an elaborate stepwise fashion to a number of fields of graduate endeavor (*15*).

This differentiation would have little effect on women's chances to enter graduate school if it were unrelated to difficulty of entry. But it is not. Although it would appear in a logical sense that the departments requiring more mathematics would be more difficult to enter, in fact it appears to be those requiring less mathematics that are the more difficult. (For the 83 graduate programs with matching undergraduate majors, the Pearson r between proportion of applicants admitted and number of recommended or required undergraduate units in mathematics or statistics is .38.) In part this may be because departments requiring less mathematics receive applications from persons who might have preferred to enter others but cannot for lack of mathematical (or similar) background, as well as from persons intrinsically inclined toward nonmathematical subjects. In part it is because in the nonmathematical subjects (that is, the humanities and social sciences) students take longer to get through their programs; in consequence, those departments have lower throughput and thus less room, annually, to accept new students. Just why this is so is a matter of debate and of great complexity. Some of the problem may lie in the very lack of a chain of prerequisites

such as that characterizing graduate work in, let us say, the physical sciences. Some may lie in the nature of the subject matter and the intractability of its data and the questions asked of the data. Some may lie in the less favorable career opportunities of these fields and in consequence a lower pull from the professional employment market. Some may lie just in the higher proportion of women enrolled and the possibility that women are under less pressure to complete their studies (having alternative options of social roles not open generally to men) and have less favorable employment possibilities if they do complete, so that the pull of the market is less for them. Whatever the reasons, the lower productivity of these fields is a fact, and it crowds the departments in them and makes them more difficult to enter.

The absence of a demonstrable bias in the graduate admissions system does not give grounds for concluding that there must be no bias anywhere else in the educational process or in its culmination in professional activity. Our intention has been to investigate the general case for bias against women in a specific matter—admission to graduate school—not only because we had the data base to do so but also because allegations of bias in the admissions process had been aired. Our approach in the beginning was naive, as befits an initial investigation. We found that even the naive question could not be answered adequately without recourse to sophisticated methodology and careful examination of underlying processes. We take this opportunity to warn all those who are concerned with problems of bias about these methodological complexities (*16*).

We also find, beyond this immediate area of concern in graduate admissions, that the questions of bias and discrimination are more subtle than one might have imagined, and we mean this in more than just the methodological sense. If prejudicial treatment is to be minimized, it must first be located accurately. We have shown that it is not characteristic of the graduate admissions process here examined (although this judgment does not eliminate the possibility of individual cases of prejudicial treatment, and it does not deal with politically or morally defined null hypotheses). The fairness of the faculty in admissions is an important foundation for further effort. That effort can be made directly by universities in seeking to equalize the progress of men and women toward their degrees (*17*). A university can use its powers of suasion to equalize the preparation of girls and boys in the primary and secondary schools for entry into all academic fields. By its own objective research it may be able to determine where and how much bias and discrimination exist and what the suitable corrective measures may be.

Summary

Examination of aggregate data on graduate admissions to the University of California, Berkeley, for fall 1973 shows a clear but misleading pattern of bias against female applicants. Examination of the disaggregated data reveals few decision-making units that show statistically significant departures from expected frequencies of female admissions, and about as many units appear to favor women as to favor men. If the data are properly pooled, taking into account the autonomy of departmental decision making, thus correcting for the tendency of women to apply to graduate departments that are more difficult for applicants of either sex to enter, there is a small but statistically significant bias in favor of women. The graduate departments that are easier to enter tend to be those that require more mathematics in the undergraduate preparatory curriculum. The bias in the aggregated data stems not from any pattern of discrimination on the part of admissions committees, which seem quite fair on the whole, but apparently from prior screening at earlier levels of the educational system. Women are shunted by their socialization and education toward fields of graduate study that are generally more crowded, less productive of completed degrees, and less well funded, and that frequently offer poorer professional employment prospects.

References and Notes

1. C. R. Blyth, *J. Am. Stat. Assoc.* **67**, 364 (1972).
2. J. Neyman, *Lectures and Conferences on Mathematical Statistics and Probability* (U.S. Department of Agriculture Graduate School, Washington, D.C., ed. 2, 1952), p. 147.
3. R. A. Fisher, *Statistical Methods for Research Workers* (Oliver and Boyd, London, ed. 4, 1932).
4. Fisher's statistic is

$$F = -2\sum_{i=1}^{n} \ln p(T_i)$$

where $p(T_i)$ is the P value of the test statistic calculated for the i^{th} experiment (department). F is referred to the upper tail of a chi-square distribution with $2n$ degrees of freedom where n = number of experimental results to be aggregated, here 85. In our application here, T_i is the usual contingency table chi-square statistic, with P value obtained from a table of the chi-square distribution with 1 degree of freedom.
5. This method uses as a statistic

$$\sum_{i=1}^{n} \chi_i^2$$

having a chi-square distribution with d.f. = n (= 85 here). χ_i^2 is the usual χ^2 statistic in the i^{th} 2×2 table.
6. In this application of Fisher's statistic (*4*), T_i is $\pm$ the square root of the chi-square statistic with sign plus if there is an excess of men admitted and sign minus otherwise; the P value is the probability of a standard normal deviate exceeding T_i.
7. Transformation to linearity by simple changes of variable, for example to log (odds), is also not successful.
8. If π_i, p_i, and p'_i represent, respectively, the probability of applying to department i, the probability of being admitted given that application is to department i, and the probability of being a male given that application is to department i, then a reasonable measure of the association of the numbers p_i, p'_i is the correlation (weighted according to the share of each major in the applicant pool)

$$\rho = \Sigma\pi_i(p_i - p_\cdot)(p'_i - p'_\cdot) / [\Sigma\pi_i(p_i - p_\cdot)^2\Sigma\pi_i(p'_i - p'_\cdot)^2]^{1/2}$$

where $p_\cdot$, $p'_\cdot$ are defined by $\Sigma\pi_i p_i$, $\Sigma\pi_i p'_i$, respectively. As usual, $|\rho| = 1$ indicates linear dependence between the p_i, p'_i while $\rho = 0$ suggests "no relation." Positive values indicate "positive association" and so on.

This correlation can be estimated by substituting the observed proportions of applicants to department i, admitted applicants to department i among applicants to department i, and male applicants to department i among applicants to department i for π_i, p_i and p'_i, respectively. This is the statistic we call $\hat{\rho}$.

We can use $\hat{\rho}$ as a test statistic for the hypothesis that $\rho = 0$. To do so we need the distribution of $\hat{\rho}$ under that hypothesis. It turns out that $\hat{\rho}/\text{Var}\ \hat{\rho}^{1/2}$ has approximately a standard normal distribution. The expression Var $\hat{\rho}$ is complicated because of the statistical dependence between p_i and p'_i. Editorial considerations have prompted its deletion. It is obtainable from the authors.
9. The probability that an observation as extreme as (or more extreme than) the most extreme one would occur by chance alone, where n = number of simultaneous independent experiments or observations, and p = probability of occurrence by chance of the most extreme observation if it had been selected at random for a single observation, is $1 - (1 - p)^n$, and thus for p close to zero is approximately np.
10. Smallness of numbers of women applicants also invalidates the normal approximation used in the significance probabilities of approach B, but this can be remedied.
11. This may be expressed as

$$\sum_{i=1}^{85} (w_i)(p_i)$$

where w_i is the number of women applying to the i^{th} major and p_i is the probability of entry of any applicant into the i^{th} major, the latter being estimated from the number of admittees divided by the number of applicants.
12. W. G. Cochran, *Biometrics* **10**, 417 (1954).
13. N. Mantel, *J. Am. Stat. Assoc.* **58**, 690 (1963).
14. Further analysis of these data, in particular examination of individual units through time, is in progress.
15. Research currently being conducted by L. Sells at Berkeley shows how drastic this screening process is, particularly with respect to mathematics.
16. There is a real danger in naive determination of bias when the action following positive determination is punitive. On the basis of Table 1, which we have now shown to be

misleading, regulatory agencies of the federal government would have felt themselves justified in withholding substantial amounts of research funding from the university. A further danger in punitive action of this kind is that, being concentrated in the research area, which provides an important source of support for graduate students, it punishes not only male but also female students—women in areas in which women have traditionally been enrolled, such as the social sciences, and also pioneering women in the physical and biological sciences, where federal support has been more concentrated.

17. In fact, data in hand at Berkeley suggest a dramatic decrease in the early dropout rates of women and the disappearance of the differential in dropout rates of men and women. It will be several years before we will be able to judge whether this phenomenon is one of decreased or simply of delayed attrition.
18. If the same naive aggregation is carried out for the 85 departments used in most of the analysis, $N = 12{,}654$, $\chi^2 = 105.6$, d.f. $= 1$, $P = 0$.
19. The investigation was initiated by E.A.H., using data retrievable from a computerized system developed by V. Aldrich. Advice on statistical procedures in the later stages of the investigation was provided by P.J.B., and programming and other computation was done by J.W.O'C.

Crisis Management: Some Opportunities

International emergency cooperation involving governments, technology, and science is now foreseeable.

Robert H. Kupperman, Richard H. Wilcox, Harvey A. Smith

Many alarming trends of our present culture share common roots. Worldwide inflation, worldwide resource shortages, extensive famine, and the inexorable quest for more deadly weapons may very well reach crisis proportions if these trends continue. They serve already as examples of national and international failures of efficient resource allocation and communications. It is important that we understand the possible future implications that these failures hold and, more important, that we develop means for dealing with them.

In discussing the crisis management demanded by such situations it is tempting to start by defining what is meant by a crisis, but this is a difficult matter. Crises are matters of degree, being emotionally linked to such subjective terms as calamity and emergency. In fact it is not necessary to define crises in order to discuss problems generally common to their management, including the paucity of accurate information, the communications difficulties that persist, and the changing character of the players as the negotiations for relief leave one or more parties dissatisfied.

In a sense, crises are unto the beholder. What is a crisis to one individual or group may not be to another. However, crises are generally distinguished from routine situations by a sense of urgency and a concern that problems will become worse in the absence of action. Vulnerability to the effects of crises lies in an inability to manage available resources in a way that will alleviate the perceived problems tolerably. Crisis management, then, requires that timely action be taken both to avoid or mitigate undesirable developments and to bring about a desirable resolution of the problems.

Crises may arise from natural causes or may be induced by human adversaries, and the nature of the management required in response differs accordingly. Thus the actions required to limit physical damage from a severe hurricane and to expedite recovery from it differ substantially from the tactics needed to minimize the economic effects of a major transportation strike and to moderate the conditions which caused it. Yet each also exhibits some characteristics of the other. For example, recovery from the devastation wrought by the hurricane's wind and floodwaters brings competition among different managers whose conceptions of recovery differ: Is the goal to reestablish the status quo, including slums, or to seize upon the opportunity for urban renewal? Similarly, a transportation strike may cause such economic chaos that the Congress—535 crisis managers—might threaten to pass laws that are detrimental to a union leadership's prestige and control over its members.

It is useful to note the characteristics common to most crisis management. Perhaps the most frustrating is the uncertainty concerning what has happened or is likely to happen, coupled with a strong feeling of the necessity to take some action anyway "before it is too late." This leads to an emphasis on garnering information: military commanders press their intelligence staffs, and civil leaders try to get more out of their field personnel and management information systems. Unfortunately, few conventional information systems are equal to the task of covering unconventional situations, so managers in a crisis must frequently fall back upon experience, intuition, and bias to make ad hoc decisions (*1*).

The problems of uncertainty are exacerbated by the dynamic nature of many crises. Storms follow unpredictable courses; famine is affected by vagaries in the weather; terrorists perform apparently irrational acts; and foreign leaders, responding to different value systems or simply interpreting situations differently, select unexpected courses of action. Thus, with limited information and resources the manager may find it difficult just to keep up with rapid developments, let alone improve the overall picture of the situation.

During a crisis, not only does an involved manager suffer from poor information, but he has the problem of identifying the objectives he wishes to accomplish and ordering them by priority in accord with his limited re-

Dr. Kupperman is Chief Scientist and Mr. Wilcox is Chief of Military Affairs of the U.S. Arms Control and Disarmament Agency, Washington, D.C. 20451. Dr. Smith is Professor of Mathematics at Oakland University, Rochester, Michigan 48063.

here w_i is the number of women applying to the ith major and p_i is the probability of entry of any applicant nto the ith major, the latter being estimated from the number of admittees divided by the number of pplicants.

2. W. G. Cochran, *Biometrics* **10,** 417 (1954).

3. N. Mantel, *J. Am. Stat. Assoc.* **58,** 690 (1963).

4. Further analysis of these data, in particular examination of individual units through time, is in progress.

5. Research currently being conducted by L. Sells at Berkeley shows how drastic this screening process is, articularly with respect to mathematics.

6. There is a real danger in naïve determination of bias when the action following positive determination is unitive. On the basis of Table 1, which we have now shown to be misleading, regulatory agencies of the ederal government would have felt themselves justified in withholding substantial amounts of research unding from the university. A further danger in punitive action of this kind is that, being concentrated in the esearch area, which provides an important source of support for graduate students, it punishes not only male ut also female students—women in areas in which women have traditionally been enrolled, such as the ocial sciences, and also pioneering women in the physical and biological sciences, where federal support has een more concentrated.

7. In fact, data in hand at Berkeley suggest a dramatic decrease in the early dropout rates of women and he disappearance of the differential in dropout rates of men and women. It will be several years before we ill be able to judge whether this phenomenon is one of decreased or simply of delayed attrition.

8. If the same naive aggregation is carried out for the 85 departments used in most of the analysis, $N = 12{,}654$, $\chi^2 = 105.6$, d.f. $= 1$, $P = 0$.

9. The investigation was initiated by E.A.H., using data retrievable from a computerized system developed y V. Aldrich. Advice on statistical procedures in the later stages of the investigation was provided by P.J.B., nd programming and other computation was done by J.W.O'C.

NOTES[1]

Dear Peter,

I write to praise your article on sex bias in graduate admissions. Your article stands out in the literature that attempts to apply statistical methods to detect and measure discrimination; its publication is highly praiseworthy.

Yet it does not fully come to grips with the central, perhaps insoluble, problem of any study dealing with survey information alone. . . that is, of any study without the randomized assignment of a proper experiment or at least without a well understood underlying theoretical structure. That central problem is the inevitable arbitrariness of stratifications used in the analysis.

Your paper begins with a simple two-by-two table for Berkeley graduate admissions: Male–Female vs. Admit–Deny. The classical hypothesis test applied to those data (under assumptions of independence and homogeneity) comes out with apparent discrimination against women at very high levels of statistical significance. By more refined analysis, however, department by department, followed by recombination of the individual departmental results, a different picture emerges: now there is apparent discrimination in favor of women and at a less extreme level of statistical significance.

[1] The following correspondence has been slightly edited with the approval of the authors for this collection.

Yet suppose that another stratification had been used—for example, California residency vs. nonresidency, ethnic background, or age. Such a stratification might have been proposed in place of the departmental one, or in addition to it. There is no a priori reason to think that the apparent results on discrimination would remain the same under a different stratification, either by itself or intersecting the department stratification.

To illustrate the general point, your paper gives a hypothetical example with two departments for which proportions of acceptance for men and women are the same within each department, but where pooling the two departments gives a very different picture of apparent discrimination. (This is a standard sort of statistical example that resembles models used in factor analysis, latent structure analysis, and elsewhere. . .mixtures with stochastic independence in the separate components.)

Yet it is not clear what our conclusions would be if another dichotomy—say California residency vs. nonresidency—were considered. Suppose, to take an extreme case, that there were sharp discrimination *against* California women residents and *in favor* of nonCalifornia women. Then we might have the following frequencies (where M = Men, W = Women, A = Admit, D = Deny):

Dept. A.

	Resident		Nonresident			Total over residency		
	A	D		A	D		A	D
M	200	0	M	0	200	M	200	200
W	0	100	W	100	0	W	100	200

Dept. B.

	Resident		Nonresident			Total over residency		
	A	D		A	D		A	D
M	50	0	M	0	100	M	50	100
W	0	300	W	150	0	W	150	300

(The two tables on the right are those given in your paper.) If now we add over the two departments, we obtain

	Resident		Nonresident			Total over residency		
	A	D		A	D		A	D
M	250	0	M	0	300	M	250	300
W	0	400	W	250	0	W	250	400

and we see plainly the sharp discriminations in opposite directions for the two residency categories.

In practice one would hardly expect anything so extreme, and in fact one can manufacture cases in which the apparent picture oscillates as fresh stratifications are brought in.

This problem is a general one of inference and of practice; it arises, for example, if one thinks about what criteria to use in setting insurance premiums.

You and your colleagues would presumably argue for the primacy of the departmental stratification because the departments are the primary decision-making units. If so, that would seem to me a non sequitur, for the central difficulty is the possible presence of unsuspected or not-allowed-for causal variables.

Quite aside from the problem of stratification one might well be concerned about treating student admissions data of this sort as if they came from a probability sample of any kind. Surely heavy selective pressures and complex dependencies must exist at the stage of application making by potential students.

For the important social problem of discrimination, it seems to me too crude to use statistical data of the kind your paper discusses. Rather, the more appropriate techniques seem to me those of careful opinion elicitation and of social psychological experimentation. Both of these rely on statistical method, but they also rely on long-studied disciplines of human behavior and of psychological experimentation.

Sincerely yours,
William H. Kruskal

Dear Bill:

(1) You are of course right to raise the fundamental point of choice of stratification. I see a nonstatistical question here. What do we mean by bias? Does discrimination by a decision-making unit based on arbitrary criteria coupled with sex constitute bias if averaging on these other criteria leaves us with independence? Somehow I feel that averaging over things like residency, which are criteria considered by a decision maker, is different from averaging over departments which are the decision-making units.

(2) Your point on the nature of the data is incontrovertible. We certainly thought only vaguely of populations of putative applicants and the P values really represent only subjective measures of the strength of the evidence.

(3) Largely through ignorance and prejudice, I wonder how effective careful opinion solicitation and the techniques of social psychology would be here. I have a high regard for the powers of deception of my colleagues if they wish to deceive. Also establishing the connection between differential attitudes to the sexes and actions on admission would still rest with data such as ours.

Having said all this, I agree wholeheartedly with your main point as expressed in your second paragraph. Data of this kind cannot settle the main questions unless we make uncheckable and possibly unwarranted assumptions. What do you think of the following semifrivolous proposal for getting a more satisfactory assessment? Let's treat the applicants to each department each year as separate populations. Suppose (for large departments) before any applications are processed we select substantial subsamples of males and females and switch sexes carefully throughout the dossier before sending them on to the departments. Significant differences between the admission pattern for the sex change group and the rest of the population would be evidence of bias (in our sense). Aggregate measures of bias could then be computed in the usual way. Of course, things like your residence-sex interaction could still not be spotted.

All the best,
Peter J. Bickel

Dear Peter:

Proper experiments of the kind you describe have indeed been done, although I do not have citations at my fingertips. One article I've read describes such an experiment for college admission, and Betty Scott has told me of a similar experiment in connection with hiring young faculty. As you suggest, the ethical and practical difficulties are large, but the latter at least are not insurmountable.

That the decisions about graduate admissions lie primarily with the departments does not seem to me a compelling reason to stratify by department and then stop. First, some important decisions are not made by the department and are highly diffuse . . . decisions about where to live, whether or not to apply, where to apply, in what field, etc. Second, the node of immediate decision need not be related to the relevant causal mechanism. For example, if I take the wrong turn on an auto trip at an intersection because a sign post is twisted, the decision was mine but the cause wholly separate.

Cordially,
William H. Kruskal