An Uncertainty Principle in Demography and the Unisex Issue

JOEL E. COHEN*

The crude death rate of country A may be less than that of country B even if every age-specific death rate of country A is greater than each corresponding one of country B. This is an example of what statisticians (unjustly) call Simpson's paradox. What holds for death rates holds equally for all other demographic rates. Simpson's paradox can recur, reversing an inequality of rates, whenever an additional variable is introduced into a stratification. Repeated stratification of a finite population (e.g., by age, sex, education, income, region) may eventually produce comparison groups that are too small for a given difference in mortality to be detected. The trade-off between the increased homogeneity of highly stratified comparison groups and the decreased ability to detect small differences in probabilities of death is described here quantitatively by an uncertainty principle, which takes the form of an inequality. The possibility of encountering Simpson's paradox suggests that since sex is only one of many possible stratifying variables that appear to affect mortality, the use of mortality tables distinguished by sex and by no other variables is, in the absence of information about the importance of other variables, demographically arbitrary.

KEY WORDS: Heterogeneity; Stratification; Simpson's paradox; Sex differences in mortality; Spurious correlation; Pooling.

1. INTRODUCTION

The purpose of this article is to show that if many characteristics affect the mortality of individuals, there are intrinsic limits to the ability of demographers to answer two elementary questions:

1. In the last year, was the force of mortality less severe in country A or in country B?
2. In the last year, would my chances of surviving have been better in country A or in country B?

The arguments to be given for death rates apply equally to all demographic crude rates. Death rates are used here because they are concrete and practically important.

For simplicity assume that the vital statistical systems of countries A and B are both perfect: thus every death is recorded. Assume also that perfect censuses of both countries were completed at the beginning of the last year, so there is no uncertainty about the populations at risk of mortality. Finally, assume that information about the characteristics of those who remain alive and those who die is as detailed as desired. Their exact ages, heights, weights, medical histories, smoking and drinking habits, driving habits, and any other desired features are assumed to be known. These assumptions avoid the necessity of discussing errors or incompleteness in data.

Even in this statistical utopia, question 1 appears easier to answer than question 2. Statistics pertain to aggregates, and question 1 is a question about aggregates. To measure the mortal risks of an individual might strain any statistical system.

Unfortunately, the comfort offered by statistical aggregates is limited. Intrinsic constraints, to be derived, limit the possibility of simultaneously specifying a large number of characteristics that affect mortality and detecting a small difference in mortality between two subgroups that are matched on these characteristics. These intrinsic constraints may be stated as an uncertainty principle. Question 1 reduces to many simultaneous versions of question 2 when so many characteristics affect mortality that the subgroups specified by those characteristics reduce to single persons.

The uncertainty principle arises because the stratification of two populations, that is, the division of each population into apparently more homogeneous subgroups for purposes of comparison, may have two effects. First, stratification may reverse the apparent rank ordering of the forces of mortality affecting the two populations. This phenomenon has been known for at least 50 years (Cohen and Nagel 1934) and is familiar to most statisticians as Simpson's (1951) paradox. [Yule (1903) pointed out that if two attributes are not associated in each of two strata, then pooling the strata can sometimes produce an artifactual association of the attributes in the aggregate. He did not discuss the possibility of apparently reversing the direction of an association by pooling strata.] Second, stratification usually reduces the size of the comparison subgroups. Small comparison groups limit the possibility of deciding whether a difference between groups is real or due to random fluctuations. This fact is also well known. What is new here is the combination of these two familiar facts in a quantitative inequality that governs the resolution of comparisons of demographic rates (or conditional probabilities or any other weighted means) in two populations.

The argument has implications for the "unisex issue" lately before American courts (U.S. Supreme Court 1983) and the Congress (U.S. Congress, 1983). Stratification of a population's mortality experience by only the sex of individuals is arbitrary if other characteristics affect the force of mortality more.

*Joel E. Cohen is Professor of Populations at Rockefeller University, 1230 York Avenue, New York, NY 10021-6399. Previous versions of this work were presented in the 1982 Alfred E. Mirsky Christmas Lectures in Science at Rockefeller University, at the Office of Population Research, Princeton University (May 13, 1983), and at the 1985 annual meeting of the Population Association of America. H. Braun, H. Caswell, J. Cole, R. Freedman, A. Gibbard, N. Goldman, N. Keyfitz, G. G. Koch, W. H. Kruskal, G. Lord, J. Menken, C. F. Mosteller, N. Ryder, C. M. Sheahan, B. H. Singer, J. Szalus, C. Westoff, C. Youtz, H. Zuckerman, and two referees provided helpful background, suggestions, and detailed criticisms. This work was supported in part by U.S. National Science Foundation Grants DEB80-11026 and BSR84-07461 to Rockefeller University, the John D. and Catherine T. MacArthur Foundation Fellowship Program, and the hospitality of Mr. and Mrs. William T. Golden.

2. A HYPOTHETICAL EXAMPLE OF SIMPSON'S PARADOX

Mortality is measured by death rates. A death rate is defined as the number of deaths in a specified population in one time unit (one year) divided by the number of person-years at risk of death. The number of person-years at risk of death is often approximated by the total living midyear population. Such an approximation will suffice here.

The crude death rate is defined as the total number of deaths in a year divided by the total midyear population. An age-specific death rate is defined as the number of deaths in a particular age group divided by the midyear number of individuals in that age group. Obviously an age-specific death rate may be viewed as a crude death rate of the population consisting only of the specified age group.

Intuitively, it might appear that if the crude death rate of country A is less than that of country B, then there must exist at least one age group such that its age-specific death rate in country A is less than its corresponding rate in country B. This intuition is false.

The intuition is false because a crude death rate is a weighted average of age-specific rates, where the weights reflect the age structures (or proportions of people in each age group) of each country. If the two countries have very different age structures, the crude death rate of country A may be less than that of country B, even though every age-specific death rate in country A is higher than the corresponding rate in country B.

Here is a hypothetical example. Suppose that countries A and B each have 100 people at risk of mortality. Suppose that there are only two age groups, "young" and "old." Suppose that the population at risk and the deaths are distributed by age in each country as in Table 1.

For both young and old, the age-specific death rates are higher in country A than in country B: $25/90 > 10/40$ and $4/10 > 20/60$. Yet the crude death rate of country A, namely 29/100, is lower than the crude death rate of country B, namely 30/100. In both countries the young have a lower death rate than the old in the same country and in the other country. Country A has a much larger fraction of its population young, whereas country B has a much larger fraction of its population old.

Simpson (1951) observed the apparent paradox to which his name is attached in contingency tables (see also Blyth 1972). Lindley and Novick (1981, app. 1) derived a necessary condition for the occurrence of Simpson's paradox in a population stratified into two subgroups. Independently, Ijiri (Sunder 1983, app.) derived the same necessary condition and showed that it is sufficient. Shapiro (1982) and Paik (1985) gave graphical representations of Simpson's paradox in this situation. These studies make it clear that Simpson's paradox can occur in any comparisons of probabilities, rates, or measurements that are weighted averages of component probabilities, rates, or measurements from subgroups. Thus Simpson's paradox is a potentially widespread phenomenon that is familiar to many demographic and statistical specialists but possibly not to other scientists or the general public. The frequency with which Simpson's paradox actually occurs in real data seems never to have been studied empirically.

Table 1. Simpson's Paradox in a Hypothetical Comparison of Death Rates

	Country A			Country B		
	At risk	Deaths	Death rate	At risk	Deaths	Death rate
Young	90	25	25/90	40	10	10/40
Old	10	4	4/10	60	20	20/60
Total	100	29	29/100	100	30	30/100

As a comfort to the intuition, note that if two countries have identical age structures and the age-specific death rates in the first all exceed the corresponding rates in the second, then the crude death rate of the first must exceed the crude death rate of the second. Simpson's paradox also cannot occur in a comparison of two stationary populations or in a comparison of two stable populations differing only by a so-called neutral change in mortality (see the Appendix).

3. REAL EXAMPLES OF SIMPSON'S PARADOX

Here are some examples of Simpson's paradox in real data from demography and business.

According to Keyfitz and Flieger (1968, pp. 94, 506), every age-specific female death rate (${}_nM_x$) was larger in Costa Rica in 1960 than the corresponding rate in Sweden in 1958–1962. Figure 1 shows the two sets of death rates up to age 70. Yet the Costa Rican female crude death rate of 8.12 per 1,000 was less than the Swedish rate of 9.29 per 1,000. (In the male population, the Costa Rican age-specific rates are higher at every age except in the 75 to 79-

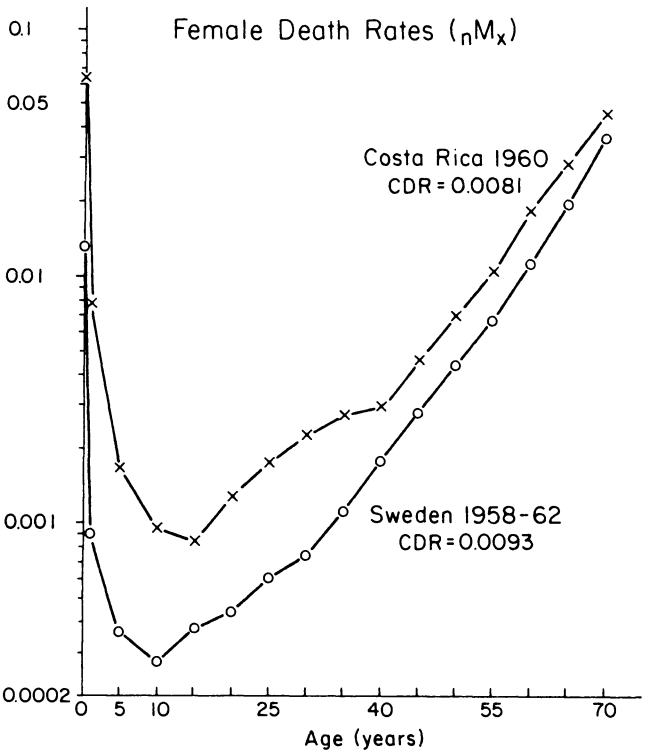

Figure 1. Age-Specific Female Death Rates of Costa Rica in 1960 and Sweden in 1958–1962. Source of data: Keyfitz and Flieger (1968). Though every age-specific death rate of Sweden is lower than the corresponding age-specific death rate of Costa Rica, the crude death rate (CDR) of Sweden exceeds that of Costa Rica.

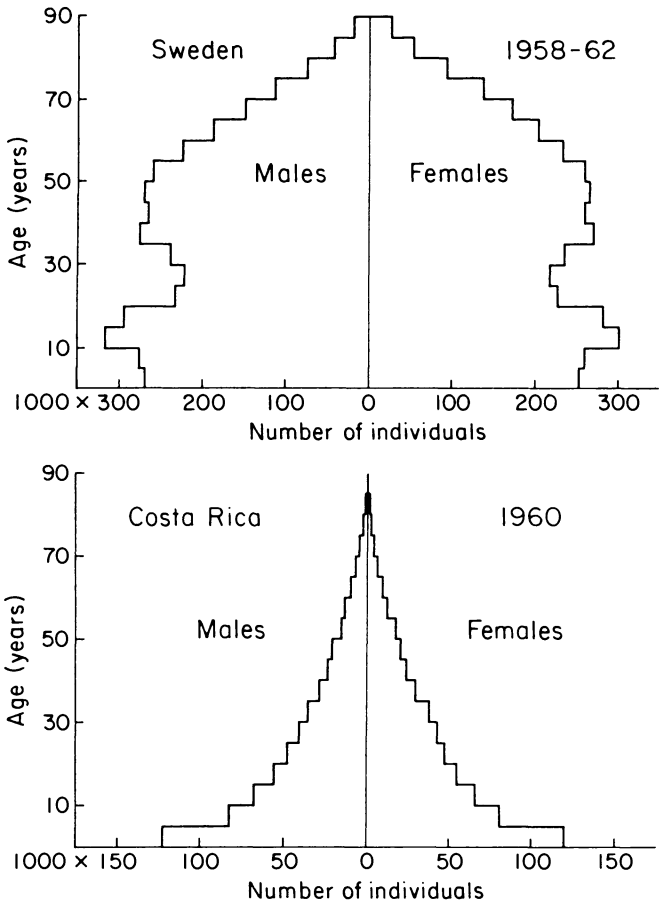

Figure 2. The Age Pyramids of Sweden in 1958–1962 and Costa Rica in 1960. Source of data: Keyfitz and Flieger (1968). The Costa Rican population has a much higher proportion of young individuals, whose death rates are less than those of old individuals in either Costa Rica or Sweden.

year-old group. Here the possibility of imperfections in the data should not be overlooked. For males also, the crude death rate of 9.15 per 1,000 is lower in Costa Rica than the rate of 10.36 per 1,000 in Sweden.)

The age pyramids of Costa Rica in 1960 and of Sweden in 1958–1962 (Fig. 2) show that Costa Rica had a much younger population than did Sweden. Because there were relatively many more Costa Ricans at the young ages for which age-specific death rates were lower than those of older Swedes, and relatively many more Swedes at the older ages for which age-specific death rates were higher than those of young Costa Ricans, the male and female crude death rates of Costa Rica were less than those of Sweden.

Real examples of Simpson's paradox were known long before Simpson (1951) attracted the attention of statisticians to the problem. Cohen and Nagel (1934, p. 449) pointed out that in 1910 the death rates from tuberculosis were higher in New York than in Richmond, Virginia, for both the "white" and the "colored" populations separately but that the aggregate death rate from tuberculosis was much higher in Richmond.

Demographers are aware of the limitations of using the crude death rate for mortality comparisons (Kitagawa 1955; Shryock and Siegel 1973, p. 418). Rural fertility and urban fertility can both be rising while (as a result of population movements) aggregate fertility is falling. The morbidity of both young and old can be improving while (as a result of shifts in the age structure) aggregate morbidity worsens.

In an example given me by Keyfitz, the mean numbers of children in 1971–1976 of French and English speakers in Quebec were, respectively, 1.80 and 1.64 and in the remaining provinces of Canada, respectively, 2.14 and 1.97, so the French exceeded the English in both comparisons. In Canada as a whole, however, the mean numbers of children of French and English speakers were, respectively, 1.85 and 1.95, so the English exceeded the French (Lachapelle and Henripin 1982).

Bickel et al. (1975) showed that among applicants for graduate school at the University of California at Berkeley, the proportion of women who were denied admission was higher than the proportion of men who were denied admission. When admissions were analyzed department by department, an apparent discrimination in favor of women appeared. The apparent overall discrimination against women resulted from the lower admission rate, for both men and women, of those departments that had more female applicants.

Using techniques for analyzing unobserved heterogeneity proposed by Heckman and Singer (1982a,b), Trussell and Richards (1985) modeled child mortality in Korea using a Weibull hazard function, which is a particular formula commonly used to model the age-specific risk of death. They found that the hazard function declined with increasing age when their analysis ignored possible heterogeneity in risks of death among children of the same age but increased with increasing age when, using the Heckman–Singer approach, their analysis allowed for possible heterogeneity. This finding illustrates that Simpson's paradox, described earlier in a simple situation, may generalize extensively. Other counterintuitive demographic consequences of unobserved heterogeneity in death rates were described by Vaupel and Yashin (1985).

Business statistics provide some interesting examples of Simpson's paradox. The overall subscription renewal rate of a magazine increased from one month to the next, but the renewal rate in each of five categories of subscriptions declined (Wagner 1982). The federal income tax rate for taxable income tax returns in each of five categories of adjusted gross income declined from 1974 to 1978, but (because of category creep) the overall tax rate increased (Wagner 1982). In the Stalcup Paper Cup case of the Harvard Business School, unit costs for each of two products increased from one period to the next when indirect costs were allocated on the basis of direct labor dollars; but when indirect costs were allocated on the basis of the weight of each product, the unit costs of both products decreased (Sunder 1983).

4. AN UNCERTAINTY PRINCIPLE FOR DEMOGRAPHIC COMPARISONS

Kruskal (1977) and Lindley and Novick (1981, p. 53) noticed that Simpson's paradox can apply recursively and that the direction of the association (e.g., of sex with probability of death, recovery, or admission) evident at the most refined level of analysis may depend on which, and how many, variables are chosen for stratification. Dawid (1979)

further pointed out that the assumption that only sex affects the probability of interest "could be examined by introducing more covariates, but at some state this process must stop, leaving a weak link at the very beginning of the chain of inference, which can only be reinforced by the statistician's informed judgement" (p. 7).

The uncertainty principle I now present, arrived at independently of Kruskal, Lindley and Novick, and Dawid, adds to their insights one elementary quantitative observation: when, as a result of repeated stratification, the groups being compared become too small, possible differences between them in the probability, rate, or measurement of interest may be swamped by statistical fluctuations.

First, I give a concrete example. Though the female crude death rate of Costa Rica is less than that of Sweden, the death rate of women aged 50–54 in Costa Rica exceeds that in Sweden. But the age-specific death rate of women aged 50–54 may be viewed as the crude rate for that particular age group. Suppose that those women were stratified according to their smoking habits into those who had never smoked tobacco and those who had. The available data do not exclude the possibility that within each smoking category, the age- and smoking-specific death rates were lower in Costa Rica than in Sweden. Then stratify each age-smoking group into two alcohol-use groups, normal and excessive. Though age- and smoking-specific rates might be lower in Costa Rica than in Sweden, age-, smoking-, and alcohol-specific rates might be lower in Sweden than in Costa Rica, and so on. With each successive stratification of an existing classification by an additional variable that affects mortality, the mortality comparison between two countries may shift from favoring one country to the other and then shift back again with further stratification.

Refinements in comparing the mortality rates of two countries must end when the numbers of people in the individual cells being compared are so small that substantively important differences in mortality rates cannot be detected with acceptable power.

To avoid being fooled by unrecognized confounding variables, one should control the comparisons between countries A and B by matching as many of the characteristics of the comparison groups as possible. (Randomization is not an option available to demographers.) The greater the number of characteristics specified, the smaller the sizes of the comparison groups.

To avoid missing a difference in mortality between groups, one should make the test for differences as powerful as possible. Given a fixed difference in the underlying force of mortality, the probability of detecting that difference increases with the sizes of the comparison groups. The greater the power desired, the greater the necessary size of the comparison groups.

I now describe quantitatively the constraints governing these conflicting goals under some idealized assumptions. Suppose that countries A and B each have populations of size N. Since Simpson's paradox requires different proportions of subgroups in the two populations being compared, suppose that each mortality-related characteristic that is used to specify comparison groups divides each country's population into two unequal groups. For simplicity, suppose that some characteristic (e.g., smoking) divides population A in the proportions 1:3 and population B in the proportions 3:1 and another characteristic (e.g., regular exercise) divides population A in the proportions 3:1 and population B in the proportions 1:3. Let C be the number of characteristics we consider, and (for convenience later) take C to be even. Suppose that within each population all C characteristics are mutually orthogonal, that is, that within any subgroup defined by some previously given characteristics, an additional characteristic splits the subgroup in the proportions 1:3 or 3:1. Let m be the number of individuals in each of two comparison groups that are specified by an even number C of characteristics, and let half of these characteristics, namely $C/2$ of them, divide population A in the proportions 1:3 and the other $C/2$ characteristics divide population A in the proportions 3:1. Then $m = N[(\frac{3}{4})(\frac{1}{4})]^{C/2}$.

Instead of using death rates, it is convenient for statistical purposes to use probabilities of death, which are estimated as the numbers of deaths during a year divided by the population at risk of death at the beginning of the year. Let p_i be the probability of death in country i, $i = 1$ (A), 2 (B), and let $q_i = 1 - p_i$. To detect a difference $d = p_2 - p_1$ in the probability of death using a two-tailed significance test at the a level with probability P, the minimum number n of individuals in each of two comparison groups (e.g., see Snedecor and Cochran 1967, p. 222) is

$$n = K(p_1q_1 + p_2q_2)d^{-2},$$

where $K = (Z_a + Z_{2(1-P)})^2$ and the probability that the absolute value of a unit normal variate will be greater than Z_a is a.

Then the maximum number C of characteristics that can be used to specify two comparison groups and the minimum difference d in probabilities of death that can be detected are constrained by the inequality

$$m \geq n \quad \text{or} \quad N(\tfrac{3}{16})^{C/2} \geq K(p_1q_1 + p_2q_2)/d^2. \qquad (1)$$

When the minimum difference d is small, the size n of each comparison group must be large, so the number C of characteristics used to stratify cannot be large. Conversely when C is large, m and therefore n must be small, so d must be large. [There are various other tests, besides the one in Snedecor and Cochran (1967, p. 222), for detecting a difference in proportions, but none of them can escape an inequality analogous to (1).]

To give a numerical illustration of (1), let $N = 1{,}200{,}000$, the approximate population of Costa Rica in 1960. Suppose that mortality is believed to be affected by eight dichotomous characteristics (sex, rural or urban location, smoking habits, drinking habits, income, education, marital status, and weight). Then the size of each comparison group is 1,483 people. Suppose that I want to have a probability of .9 of detecting a difference in either direction between two comparison groups and I want the difference to be significant at the 1% level. Then $K = 14.88$ (Snedecor and Cochran 1967, p. 113). For an annual probability of death $p_1 = .01$, which is typical of middle age, the minimum detectable difference would be $d < -.01$ or $d > .02$. Thus the annual probability of death in the comparison group would have to be $p_2 < 0$, which is impossible, or $p_2 > .03$.

In Costa Rica in 1960, the annual probability of death for males aged 30–34 was .009893, and the corresponding rate in Sweden in 1958–1962 was .006525, according to Keyfitz and Flieger (1968). In this hypothetical example, if two groups of 1,483 people each were repeatedly drawn, one from a population with the Costa Rican probability of death and one from a population with the Swedish probability of death, the difference in probability of death would not be detectable by a significance test at the 1% level in 90% of the samples. Similarly, the annual probability of death for Costa Rican females aged 30–34 was .011445, and the corresponding Swedish annual probability of death was .003676. This difference is also too small to be detected by a significance test at the 1% level in 90% of repeated samples, with each comparison group of size 1,483.

The unrealistic assumptions used to derive the preceding inequality (e.g., that each stratification orthogonally splits the population in the proportions 1:3 or 3:1 and that each comparison group in populations A and B is of the same size) can easily be removed. At the cost of more complicated expressions, an uncertainty principle appropriate to a particular real comparison can be derived using the same ideas. What matters is not the specific formulas in (1) but the general conflict between the control and the power of comparisons that the inequality illustrates.

The inequality (1) constrains inferences about a single pair of matched comparison groups. It limits the precision of answers to the second elementary question posed in the Introduction. To compare a given individual's mortality risks in country A and country B, one seeks a single pair of comparison groups that share as many as possible of that individual's characteristics.

To answer the first elementary question posed in the Introduction, that is, to compare the mortality risks of country A and country B, information from all strata should be used. Various summary statistics or procedures for simultaneous inference are possible. To take one illustrative example, suppose that in the jth stratum, $j = 1, 2, \ldots, 2^C$, the significance level of a test for a difference between the two comparison groups in the probability of death is α_j. Assuming independence among strata in mortality differences, the summary statistic $q = -2\Sigma_j \ln \alpha_j$, originally proposed by Fisher, has the distribution of χ^2 with 2^{C+1} df. Gill (1978, pp. 75–76) derived Fisher's test statistic and reviewed subsequent analysis of it as well as alternative procedures. He observed that "if the average of the α_i exceeds approximately 0.3, then combined significance [a value of q with a tail probability much less than 0.3] cannot be achieved . . . , i.e., strong evidence cannot be obtained by combining bits of rather weak evidence" (p. 76). If stratification reduces the sizes of comparison groups far enough and the differences between countries within strata are small enough, even simultaneous test procedures will not recover significant evidence of a difference between countries in the probability of death.

Other approaches to comparing the mortality of two populations do not escape from analogous uncertainty constraints. For example, if contingency tables are used, an increase in the number of dimensions implies a decrease in the number of individuals counted within each cell of the table. If proportional hazard models are used, increasing the number and refinement of covariates decreases the possibility of comparing the assumed underlying common hazard function with that of any of the subgroups specified by the covariates, because the subgroups become too small. Related uncertainty principles for more complicated models may be developed by using the approach of Heckman and Singer (1982a,b).

5. THE UNISEX ISSUE

Comparing the mortality of the populations of two countries does not differ in substance from comparing the mortality of any two arbitrarily defined populations. The preceding argument shows that the subgroup-specific and overall mortality risks of population A cannot reliably be distinguished from those of population B if so many stratifying variables affect mortality that too few individuals belong to the subgroups being compared and the probabilities of death in the subgroups differ between A and B by too little. In this case, whether A or B is seen as having better mortality risks could depend in an arbitrary way on how many stratifying variables are specified.

We now leave the realm of elementary calculations, where facts are clear, to enter the realm of interpretation, where reasonable people may differ.

According to the Teachers Insurance and Annuity Association—College Retirement Equities Fund (TIAA–CREF; 1983), "The [unisex] issue is whether the use of mortality tables that reflect the differences in life expectancy between men and women will be prohibited, required, or left as one of several approaches to determining annuity income under pension and TDA [tax-deferred annuity] plans" (p. 7).

A background press release (TIAA–CREF 1980) explained:

> At issue in the controversy is whether Title VII of the Civil Rights Act of 1964 and/or the Equal Pay Act are violated by the use of "sex-distinct" mortality tables that reflect known differences in the life expectancies of men and women. One result of using such tables is that women receive somewhat lower monthly pension benefits than similarly situated men under single-life retirement income options, because women live longer on average and therefore receive more monthly payments than men of the same age. (p. 2)

The U.S. Supreme Court (1983) ruled that Title VII of the Equal Rights Act of 1964 forbids employers from offering as a privilege of employment any retirement annuity plan, whether operated by the employer or by a third party under contract with the employer, that does not give men and women equal monthly retirement payments for equal contributions made after August 1, 1983. The Court noted:

> No insurance company has been joined as a defendant, and our judgment will in no way preclude any insurance company from offering annuity benefits that are calculated on the basis of sex-segregated actuarial tables. All that is at issue in this case is *an employment practice*: the practice of offering a male employee the opportunity to obtain greater monthly annuity benefits than could be obtained by a similarly situated female employee. (p. 14, footnote 17)

The Court did not rule out employees' accumulating tax-deferred withholdings from income under an employer's plan, taking a lump-sum payment on retirement, and buying a sex-segregated annuity on the open market.

The Court had ruled in 1978 that when pension benefits are set at the same level for men and women in a retirement plan, an employer may not require larger contributions from women. This earlier ruling, which governs so-called "defined benefit plans," and the 1983 ruling, which governs "defined contribution plans," together settled the unisex issue for employers.

In 1983 the U.S. House of Representatives and Senate considered (e.g., U.S. Congress 1983) bills "to prohibit discrimination in insurance on the basis of race, color, religion, sex, or national origin" (H. R. 100) and "to promote interstate commerce by prohibiting discrimination in the writing and selling of insurance contracts" (S. 372). The bills did not pass and as of May 16, 1985, were not before the Congress. Is additional legislation to prevent insurers from selling annuities based on sex-segregated actuarial tables reasonable or desirable?

The Fourteenth Amendment of the U.S. Constitution, in guaranteeing to all persons equal protection of the laws, has been interpreted to require "that those who are similarly situated be similarly treated" (Tussman and tenBroek 1949, p. 344). Here *similarly situated* is to be defined with respect to the purpose of the law (Tussman and tenBroek 1949). If new laws are intended to assure "fairness" in the criteria according to which the prices of insurance are set, then the acceptability of criteria (such as sex, race, or age) is largely influenced by social and political judgments of what is "fair." But these judgments seem subject to influence by scientific and statistical findings. For example, if as has been claimed, nonsmoking American men and women have identical life expectancies (apart from deaths from accidents, suicide, and homicide), it would seem very difficult to defend charging male and female nonsmokers different rates because more men than women smoke. In the other direction, the absence of information is less persuasive: judgments of fairness might forbid insurance premium rates that differ by sex even if extensive stratification on other variables failed to explain sex differences in mortality.

It has been objected that if insurers are not permitted to use sex-distinct mortality tables, why should they be permitted to use age-specific mortality tables? According to this argument, both sex and age are biologically determined variables that have measurable effects on mortality, and since it seems absurd to disregard age, it would seem equally absurd to disregard sex.

One response to this argument is that insurers sometimes ignore age or use extremely broad age classes in setting premiums, for example, when health insurance premiums are set on the basis of prior health history, or increase with age only for individuals over 50. This shows that under certain circumstances, age is not a necessary variable. Even if sex is like age, sex need not necessarily be considered.

A second response to this argument is that age is not like sex because an individual's age changes over time, whereas sex does not (ordinarily).

Another issue that has been proposed as relevant to the unisex issue is whether the observed mortality differences between males and females are biologically determined or are consequences of cultural, social, and behavioral traits for which biological sex is a surrogate but not intrinsically responsible in some sense. Those who raise this issue argue that if the sex differences in mortality are primarily biologically determined, then sex-distinct mortality tables should be used, but if the sex differences in mortality are not primarily biologically determined, then sex-distinct mortality tables should not be used. [In most, but far from all, nonhuman animal species in which the survival of males and females has been compared, the females survive better. But the difference in survival between the sexes can be greatly influenced by changes in the environment (MacArthur and Baillie 1932; Comfort 1979, pp. 163–167).]

One problem with raising this issue of biological determination is that it asks a question that is very difficult to define precisely and nearly impossible to answer persuasively.

A second problem with this issue is that it has not been necessary to answer the analogous question regarding biological versus cultural determination of mortality differences between "races" (aside from the fact that what race a person belongs to in this country is much more a social than a biological question). A social decision has been made that race-distinct mortality tables will not be used for insurance, although the race- and sex-distinct tables published by the U.S. National Center for Health Statistics are used for a variety of other purposes. For employers, the U.S. Supreme Court (1983) ruled, "if it would be unlawful to use race-based actuarial tables, it must also be unlawful to use sex-based tables" (p. 9). For insurers, a social decision, one way or the other, could equally be reached with respect to the use of sex-distinct mortality tables without resolving the question of biological versus cultural determination.

The decision to permit or forbid sex-distinct actuarial tables, with or without other stratifying variables, could be enlightened by research in a combination of statistical demography and economics. What is needed in statistical demography is a rank ordering of variables (let us suppose dichotomous variables, for simplicity) according to their ability to discriminate mortality, conditional on the subgroupings higher in the rank ordering. For example, suppose that smoking habits (never smoked vs. ever smoked) are the best single discriminator of life expectancy at some age, that weight is the best discriminator, conditional on smoking, that sex is the best discriminator, conditional on smoking and weight, and so on. It would then be useful to compare quantitatively both the costs for insurers and insureds of measuring reliably the stratifying variables (e.g., smoking, weight, sex, etc.) and the gains (or losses) of charging more risk-specific premiums for more risk-homogeneous groups according to, for example, smoking only, or smoking and weight, or smoking, weight, and sex, and so forth. This combination of economic and demographic analysis could, in principle, give a revised order of merit and a quantitative estimate of merit for potential stratifying variables. Carrying out this program in a convincing way may not be easy (Kruskal 1984).

If it turned out that sex stood by itself at the head of the ordering of stratifying variables, not nearly approached in merit by any other variable, then it would be possible to argue for the economic and practical sense of using only sex as a stratifying variable for actuarial tables. If it turned

out that sex stood in the middle of several influential variables, all of high merit, it would be possible to argue in favor of using sex as one among several stratifying characteristics. If sex turned out to have little merit as a stratifying variable, the use of only sex in actuarial tables could hardly be defended.

According to a letter to *The Washington Post* from Congressman John D. Dingell,

> [There] is a greater mortality differential between Mormons and non-Mormons, between whites and blacks, between Jews and non-Jews at older ages, between smokers and nonsmokers, between residents of Hawaii and residents of the District of Columbia, than there is between males and females. All of these, except sex, are ignored in rate and benefit classification for annuities. (U.S. Congress, House 1983, p. 52)

This intriguing list, if confirmed, is a beginning but not a full response to the program of research just proposed.

To summarize the argument: Sex is only one of a number of characteristics of individuals and groups that appear to affect mortality and life expectancy. The decision to permit or forbid the use of sex-distinct actuarial tables in the insurance industry does not hinge on an analogy with the use of age or on whether sex differences in mortality or morbidity are biologically determined. Rather, the decision rests ultimately on contemporary values of fairness, and these values can be constructively influenced by demographic and economic analyses that do not appear to have been done yet. Such analyses could show whether sex is the sole stratifying variable of demographic and economic merit, one of several such variables, or of negligible merit as a stratifying variable.

In the absence of reasonable attempts to carry out such demographic and economic analyses, the decision to stratify by sex alone is arbitrary. The use of sex alone, apart from age, fails to recognize that an apparent advantage in mortality of one sex over the other may well be reversed or eliminated on further stratification.

APPENDIX: DEMOGRAPHIC COMPARISONS IN WHICH SIMPSON'S PARADOX CANNOT OCCUR

Robert Parke (personal communication, March 29, 1985) asked whether Simpson's paradox could occur in comparing age-specific and crude death rates of two stationary populations. It cannot. If $\mu_A(x)$, the force of mortality at age x in population A, is greater than $\mu_B(x)$ for every age x, then d_A, the crude death rate in population A, must exceed d_B.

Proof. Let $l_A(x)$ be the fraction of a birth cohort in population A that survives to age x. Thus $l_A(0) = 1$. Then $\mu_A(x) > \mu_B(x)$ for all x implies that

$$\int_0^x \mu_A(t)\,dt > \int_0^x \mu_B(t)\,dt$$

for all $x > 0$, which implies that

$$l_A(x) = \exp\left(-\int_0^x \mu_A(t)\,dt\right) < l_B(x) = \exp\left(-\int_0^x \mu_B(t)\,dt\right)$$

for all $x > 0$. Let e_A be the expectation of life at birth in population A. Then

$$e_A = \int_0^x l_A(x)\,dx < e_B = \int_0^x l_B(x)\,dx$$

so $d_A = 1/e_A > d_B = 1/e_B$. This proof shows that if $\mu_A(x) \geq \mu_B(x)$ for all x, then $d_A \geq d_B$, and if in addition $\mu_A(x) > \mu_B(x)$ over any interval of x with positive length, then $d_A > d_B$.

If the forces of mortality in stable populations A and B satisfy $\mu_A(x) = \mu_B(x) + \varepsilon$, where ε is independent of age (a "neutral" change in mortality; see Keyfitz 1968, pp. 187–188), but the schedules of fertility in A and B are identical, then again Simpson's paradox cannot arise because $d_A = d_B + \varepsilon$.

Proof. The stable fraction of the population between age x and $x + dx$ is the same in both populations, $c_A(x)\,dx = c_B(x)\,dx$ (Keyfitz 1968, p. 188). Since the crude death rate is $\int_0^\infty c(x)\mu(x)\,dx$ (Keyfitz 1968, p. 172), it follows that

$$d_A = \int_0^\infty c_A(x)\,\mu_A(x)\,dx = \int_0^\infty c_B(x)[\mu_B(x) + \varepsilon]\,dx$$

$$= \int_0^\infty c_B(x)\,\mu_B(x)\,dx + \varepsilon\int_0^\infty c_B(x)\,dx = d_B + \varepsilon.$$

[*Received September 1983. Revised May 1985.*]

REFERENCES

Bickel, P. J., Hammel, E. A., and O'Connell, J. W. (1975), "Sex Bias in Graduate Admissions: Data from Berkeley," *Science*, 187, 398–404. Reprinted in: *Statistics and Public Policy*, eds. W. B. Fairley and C. F. Mosteller, Reading, MA: Addison-Wesley, 1977, pp. 113–130.

Blyth, C. R. (1972), "On Simpson's Paradox and the Sure-Thing Principle," *Journal of the American Statistical Association*, 67, 364–366.

Cohen, M. R., and Nagel, E. (1934), *An Introduction to Logic and Scientific Method*, New York: Harcourt, Brace.

Comfort, A. (1979), *The Biology of Senescence* (3rd ed.), New York: Elsevier.

Dawid, A. P. (1979), "Conditional Independence in Statistical Theory," *Journal of the Royal Statistical Society*, Ser. B, 41, 1–31.

Gill, J. L. (1978), *Design and Analysis of Experiments in the Animal and Medical Sciences*, Ames: Iowa State University Press.

Heckman, J. J., and Singer, B. (1982a), "The Identification Problem in Econometric Models for Duration Data," in *Advances in Econometrics*, ed. W. Hildenbrand, New York: Cambridge University Press, pp. 39–77.

——— (1982b), "Population Heterogeneity in Demographic Models," in *Multidimensional Mathematical Demography*, eds. K. Land and A. Rogers, New York: Academic Press, pp. 567–599.

Keyfitz, N. (1968), *Introduction to the Mathematics of Population*, Reading, MA: Addison-Wesley.

Keyfitz, N., and Flieger, W. (1968), *World Population: An Analysis of Vital Data*, Chicago: University of Chicago Press.

Kitagawa, E. M. (1955), "Components of a Difference Between Two Rates," *Journal of the American Statistical Association*, 50, 1168–1194; Corrigenda (1956), 51, 651.

Kruskal, W. H. (1977), "Notes" [on Bickel et al. (1975)], in *Statistics and Public Policy*, eds. W. B. Fairley and C. F. Mosteller, Reading, MA: Addison-Wesley, pp. 127–129.

——— (1984), "Concepts of Relative Importance," *Queštró*, 8, 39–45.

Lachapelle, R., and Henripin, J. (1982), *The Demolinguistic Situation in Canada: Past Trends and Future Prospects*, Montreal: Institute for Research on Public Policy and Brookfield Publishing Co.

Lindley, D. V., and Novick, M. R. (1981), "The Role of Exchangeability in Inference," *Annals of Statistics*, 9, 45–58.

MacArthur, J. W., and Baillie, W. H. T. (1932), "Sex Differences in Mortality in *Abraxas*-Type Species," *Quarterly Review of Biology*, 7, 313–325.

Paik, M. (1985), "A Graphic Representation of a Three-Way Contingency Table: Simpson's Paradox and Correlation," *The American Statistician*, 39, 53–54.

Shapiro, S. H. (1982), "Collapsing Contingency Tables—A Geometric Approach," *The American Statistician*, 36, 43–46.

Shryock, H. S., Siegel, J. S., and Associates (1973), *The Methods and Materials of Demography* (rev. ed.), Washington, DC: Government Printing Office.

Simpson, E. H. (1951), "The Interpretation of Interaction in Contingency Tables," *Journal of the Royal Statistical Society*, Ser. B, 13, 238–241.

Snedecor, G. W., and Cochran, W. G. (1967), *Statistical Methods* (6th ed.), Ames: University of Iowa Press.

Sunder, S. (1983), "Simpson's Reversal Paradox and Cost Allocation," *Journal of Accounting Research*, 21, 222–233.

Teachers Insurance and Annuity Association—College Retirement Equities Fund (1980), "EEOC Approves TIAA–CREF Unisex Mortality Table for Future Pension Premiums," March press release, New York: Author.

——— (1983), *1982 Annual Report*, New York: Author.

Trussell, J., and Richards, T. (1985), "Correcting for Unmeasured Heterogeneity in Hazard Models Using the Heckman–Singer Procedure," in *Sociological Methodology*, ed. N. Tuma, San Francisco: Jossey-Bass, pp. 242–276.

Tussman, J., and tenBroek, J. (1949), "The Equal Protection of the Laws," *California Law Review*, 37, 341–381.

U.S. Congress, House (1983), Committee on Energy and Commerce, Subcommittee on Commerce, Transportation, and Tourism, Nondiscrimination in Insurance Act of 1983: Hearings on H.R. 100, February 22 and 24, 1983, 98th Cong., 1st Sess., Ser. No. 98-35, Washington, DC: Government Printing Office.

U.S. Supreme Court (1983), Arizona Governing Committee for Tax Deferred Annuity and Deferred Compensation Plans, etc., et al., Petitioners v. Nathalie Norris etc., Slip Opinion No. 82-52, Washington, DC: Reporter of Decisions, Supreme Court of the United States.

Vaupel, J. W., and Yashin, A. I. (1985), "Heterogeneity's Ruses: Some Surprising Effects of Selection on Population Dynamics," *The American Statistician*, 39, 176–185.

Wagner, C. H. (1982), "Simpson's Paradox in Real Life," *The American Statistician*, 36, 46–48.

Yule, G. U. (1903), "Notes on the Theory of Association of Attributes in Statistics," *Biometrika*, 2, 121–134.