Sociological Perspectives
Vol. 40, No. 1, pp. 129-154

ISSN 0731-1214

EFFICIENTLY STUDYING RARE EVENTS: Case-control Methods for Sociologists

MICHAEL G. LACY*
Colorado State University

ABSTRACT: *Case-control designs involve samples stratified disproportionately on a binary dependent variable. This design, though infrequently used by sociologists, offers tremendous logistical efficiency in the study of rare events, such as divorce, joining a religious cult, or committing a crime. This paper attempts to sensitize sociologists to the many situations in which this design is useful, and offers general and accessible guidance on the practice of case-control research. Using the epidemiologic literature, I explain the underlying logic of case-control design, discuss how to conduct case-control sampling, and briefly cover data analysis issues. I conclude with an empirical illustration of a case-control study examining factors associated with the change of chief administrative officers at post-secondary educational institutions.*

With the contemporary popularity of event history methods (Allison 1984; Tuma and Hannan 1984; Yamaguchi 1991), the study of binary outcomes in time (events) enjoys considerable attention among sociologists. A common difficulty in such studies is that events of interest may be rare across time and persons. The canonical example here is suicide, the rareness of which makes an individual-level longitudinal study impractical, and may have led Durkheim (1951) to use aggregate data to test an individual-level theory. Many other rare events hold theoretical and substantive interest for sociologists, including committing homicide, joining a religious cult, becoming divorced, or the demise of an organization. These events may be common in absolute numbers, but they are sparse if viewed in terms of the space of person-time (marriage-time, organization-time, etc.).[1] Consequently, studying their occurrence with conventional longitudinal

*Direct all correspondence to: Michael G. Lacy, Department of Sociology, Colorado State University, Fort Collins, CO 80523-1784; *e-mail*: mglacy@lamar.colostate.edu.

research designs is expensive and time-consuming, if possible at all. For example, the contemporary U.S. divorce rate (divorces per married woman per year) is about 0.02 (U.S. Bureau of Census 1995). A longitudinal study of the determinants of divorce therefore would require a sample of some 10,000 marriage-years to accumulate 200 new divorces. (For examples of panel studies of divorce, see, e.g., Diekmann and Klein 1991; Booth, Johnson, White, and Edwards 1991). Such problems in studying rare outcomes can occur with any unit of analysis. Suppose we wish to study the demise of colleges and universities as a means to understanding organizational adaption to environments. Because one-half percent of colleges and universities "die" per year in the United States (Peterson's Guides 1995), a sample of 20,000 institution-years would be needed to achieve an expectation of even 100 institutional deaths.

The goal of the current paper is to better acquaint the general quantitatively oriented sociological audience with a solution to the problem of studying rare events. To do so, I provide an introduction to "case-control" or "response-based" study designs (Manski 1995; Rothman 1986). These designs involve samples stratified disproportionately on the *dependent* variable. I wish to explain the validity of this design and sensitize sociologists to its potential. Therefore, my discussion emphasizes the underlying logic of case-control design and some key issues in sampling, with relatively little emphasis on data analytic techniques. An illustrative empirical example is also presented.

THE DISUSE OF "RESPONSE-BASED" OR "CASE-CONTROL" STUDIES IN SOCIOLOGY

A efficient solution to the expense and inefficiency encountered in studying rare events is known as "case-control" or "retrospective" design in epidemiology (Rothman 1986; Schlesselman 1982), and "choice-based" or "response-based" sampling in the econometric literature (Manski 1995; Manski and Lerman 1977). These designs offer tremendous efficiencies, to be described below, but have been used infrequently (and not always correctly) in sociology. Although some of the econometric literature on choice-based sampling has diffused directly into sociology (Bye, Gallicchio, and Levy 1987; Manski 1981; Xie and Manski 1989), that literature is quite technical and has focused primarily on problems in statistical estimation. The econometric literature is excellent and highly sophisticated, but has been insufficiently appreciated by most sociologists,[2] and, as compared to the epidemiologic literature, has paid less attention to some issues of sample selection and analysis that the ordinary quantitative sociologist might find useful. Therefore, my presentation of case-control design is grounded in the epidemiologic literature, as its relative accessibility and attention to sampling issues makes it potentially valuable to a wide range of sociologists.

SOME ORIENTING SOCIOLOGICAL ILLUSTRATIONS OF CASE-CONTROL STUDIES

The defining feature of a case-control or response-based study is that the sample is stratified on a discrete *dependent* variable, not, as in conventional sociological practice, on one or more *independent* variables. As an orienting illustration, let us return to thinking about divorce, and consider how a case-control study of divorce might be done. The investigator would enumerate all the couples in a small city who divorced during a one year period ("the cases"), data readily available in the divorce reports of many newspapers. For comparison, the cases would be supplemented with a sample from the population at risk—all married couples who lived in the city at some point during that time period ("the controls"). The controls might be obtained by random-digit dialing, with screening for marital status and place of residence during the time of interest. Let us presume that 200 divorces occurred during the time of interest, and that a sample of 200 controls were obtained for comparison. The composite sample would be disproportionately stratified on the dependent variable: The probability of selection would be 1.0 for couples who divorced, and perhaps 50 times lower for controls, assuming the prevailing U.S. divorce rate. Note that this study is retrospective, a typical (though not necessary) feature of case-control designs, because data was collected after the events of interest occurred, rather than through a longitudinal panel study (or "follow up" design, in the epidemiologist's parlance). Proceeding from a theory of homogamy as a factor in marital stability, the investigator might wish to estimate the effect of age differences on the risk of divorce. As will be shown below, if relatively simple though sometimes counter-intuitive details of sampling procedure and analysis are followed, the case-control study sketched here would yield a valid estimate of the relative effect of age differences on the hazard rate for divorce, just as could be obtained from an event history or similar approach applied to data from a large panel of married couples (Diekmann and Klein 1991). However, and this is a key point of interest, a researcher using a case-control design would obtain quite similar results at a small fraction of the time and money cost of a longitudinal study.

The rarer an event is, the greater such efficiency will likely be. A rural sociologist might wish to explore factors that cause farmers to convert to sustainable or "organic" agricultural practices by following a panel of farmers, year after year, until a sufficient number converts to "organic" production. Similarly, one might wish to study what affects the rate of attaining a graduate degree among persons of minority ethnic status without the expense of following a panel or cohort of college entrants. Both of these hypothetical studies could be done relatively inexpensively and quickly using case-control designs. The feasibility of such studies, and that of case-control studies in general, depends on how easy it is to obtain some sort of listing of persons (units) who experienced the rare event. Given the existence of certifying organizations for organic agricultural producers, and list-

ings of ethnic-minority PhD's, this feasibility condition should be fulfilled for both of these hypothetical studies.

THE PRE-HISTORY OF CASE-CONTROL DESIGNS IN SOCIOLOGY

Although sociologists have at least occasionally (and in most instances probably quite unconsciously) used case-control designs, such studies are infrequent though not absent from the sociological literature (See note 2). An interesting example of this pre-conscious use of a case-control design is Eitzen's (1970) study of factors affecting "strong support" for the 1968 presidential candidate George Wallace. In this study, Eitzen obtained a sample of strong Wallace supporters ("cases") by spotting automobiles with Wallace bumper stickers, obtaining owners' addresses through license plate address listings, and then screening for residence in the city. He then selected a random sample of controls from local listings of automobile registrants. Both cases and controls were interviewed concerning their political attitudes, occupations, and demographic characteristics. Eitzen was able to show a strong relationship between status inconsistency and being a Wallace supporter, substantiating certain conventional theories concerning status consistency and political extremism. Although Eitzen (personal communication, 1995) was quite unaware of case-control methods at the time, his study was in most respects relatively correct, and provides a sociologically interesting example of what the method can offer.

Had Eitzen (1970) followed the meager literature on case-control designs then available in sociology, he never would have done the study. At that time, the sole substantial commentary in sociology on case-control designs appeared in Hirschi and Selvin's (1967) methodological classic on survey analysis. In discussing correct description, prediction, and causal analysis using contingency tables, they criticize the Gluecks' (e.g., 1950) studies of delinquency, arguing that causal inference cannot be made from data comprising samples of persons who do and do not exhibit the behaviors of interest, that is, in the current language, a case-control study. Hirschi and Selvin argue that such a sample, stratified on a dependent variable representing a rare event or behavior, cannot represent the population, in which the marginals for the dependent variable must differ markedly from the sample. The overall thrust of Hirschi and Selvin's argument was that such samples, while useful for describing the conditional proportions of the independent variable within categories of the dependent variable, are useless for causal or predictive analysis unless the marginal frequencies for the dependent variable are known so as to permit weighting to adjust for stratification on the dependent variable.

Hirschi and Selvin were simply wrong, an unfortunate feature of an otherwise classic text that has enlightened generations of sociologists and which still merits reading for its advice on causal thinking in sociology. If Hirschi and Selvin had been familiar with some of the epidemiologic literature available in the 1960s (e.g., Cornfield 1951; Cornfield and Haenszel 1960), they might have thought

otherwise. That literature, its epidemiologic descendants (e.g., Mietinnen, 1976), work in the 1980s on categorical data analysis (e.g., Fienberg 1980), and the econometric literature on "choice-based" sampling (Manski and Lerman 1977; Manski and McFadden 1981) all have shown that valid estimates of relationship or effect can be obtained from retrospective, case-control sampling without knowledge of the marginals on the dependent variable, although such information can be useful (Manski 1995).

RECENT ECONOMETRIC AND OTHER SOCIAL SCIENCE LITERATURE ON "CHOICE-BASED" SAMPLING

In the 1970s, econometricians (Manski and Lerman 1977; McFadden 1973) began exploring mathematical and statistical models for "discrete choice," and part of this literature concerned what they termed "choice-based" sampling. This developed in the context of the traditional interest of economists in human choice behavior. Consider, for example, consulting local auto dealers and obtaining a sample of persons who purchased Fords, one that purchased Hondas, and one that purchased Pontiacs. Ordinary methods of estimating the effect of gender or occupation on brand choice, such as a simple comparison of percentages, will not yield valid results unless weighting is done to reflect stratification on the dependent variable. However, Manski and McFadden (1981) and others developed statistical models that permit valid estimates of effect without knowledge of the population marginals on the dependent variables. They also developed methods for using auxiliary data in place of full knowledge of the population marginals (Manski and Lerman 1977; Manski 1995). This methodological literature has continued up until the present, and has maintained a primary focus on the consistency and efficiency of statistical estimators, rather than on sampling or other design issues. As indicated previously, this literature is extremely sophisticated, and has found considerable use in related social science disciplines, such as geography (e.g., Thill and Horowitz 1991), but apparently and regrettably has received little use by sociologists, as substantive articles by sociologists rarely have cited this literature (See note 2).

In a different line of intellectual development, many of the contemporary classics of categorical variable analysis, written by statisticians or biostatisticians but widely appreciated among sociologists (Agresti 1990; Fienberg 1980), have given at least some discussion to the applicability of certain statistical models (typically log linear) to "retrospective sampling," indicating that odds-based techniques are appropriate for such samples, since they offer estimates of effect or relationship that are invariant with respect to sampling ratios on the dependent variable. Thus, implicitly, such approaches can be used to correctly analyze "choice-based" or "case-control" data, even in situations in which the marginals on the dependent variable are unknown and so do not permit weighting to reflect stratification. Again, as in the econometric literature, the primary focus is on data analysis, with little treatment of design issues. The absence of emphasis on case-control

design issues in both the econometric and categorical variable literatures is not a scholarly failing; it simply reflects that these bodies of work were not meant to offer design advice to the ordinary quantitative sociologist, nor was it their particular intent to draw the sociologist's attention to the logic and potential efficiencies of case-control designs.

Despite the relative inattention to case-control/choice-based designs in sociology, sociologists have occasionally used this design quite consciously. Gortmaker (1979) studied the effect of poverty on infant mortality using a case-control design. His article received critical methodological comment (Swafford 1980; 1981), but Gortmaker's (1981) response appears to have had the better of the debate, demonstrating a clear understanding of the epidemiologic and biostatistical literature on how to conduct, analyze, and interpret case-control studies. More recently, Arnold and Hagan (1992) published an event history analysis using a case-control design, with the event of interest being the relatively rare occurrence of attorneys being formally sanctioned for misconduct. Both of these articles handled the research design and analysis of data in a way quite in concert with the methodological literature in epidemiology and biostatistics. Their use of the case-control design, and moreover its *correct* use and analysis, still represents a relative rarity in sociology, due in part to the absence of an accessible source to guide sociologists on how to conduct and analyze such studies.

Regarding the potential, conduct, and analysis of case-control designs, the most relevant social science literature has appeared in criminology. Two articles have presented basic discussions for the quantitative criminologist on the practice (Goodman, Mercy, Layde, and Thacker 1988) and analysis (Loftin and MacDowall 1988) of case-control studies of the etiology of crime. Once again, as judged from citations in the published journal literature, their work undeservedly has received almost no attention from sociologists in general, nor even from criminologists.

Thus, despite the work of econometricians and categorical data analysts on statistical techniques for choice-based/case-control samples, the existence of at least some substantive case-control studies in sociology, and the relatively recent appearance of some literature in criminology, sociologists still have no accessible source in sociology proper to which they can turn for general guidance on case-control studies. This presumably accounts for the relative disuse of this design.

VALIDITY OF CASE CONTROL DESIGNS

Binary Dependent Variables, Events, and Rates

My explanation of the logic and validity of case-control or response-based designs rests on the methodological literature of epidemiologists, who have used case-control designs widely for over 40 years. By contrast to the econometric tradition of approaching these studies from the perspective of statistical estimation, epidemiologists have taken a different though not contradictory angle,

emphasizing issues of design and sampling. Despite their focus on biomedical variables, their literature is generally quite accessible. More importantly, the epidemiologic approach, as we shall see, directly connects with sociologists' recent interests in event history.

Epidemiologic thinking rests fundamentally on conceiving of binary dependent variables as events—things that occur to persons within the space of "person-time." Familiar examples include the occurrence of a heart attack, a cancer, or even a common cold. Even after the advent of event history methods, most sociologists do not automatically think in these terms, but they might well do so in many situations. Adopting this slight change of perspective is helpful in entering the epidemiologic literature. Thus, instead of thinking of a person *being* an organic farmer or a minority-status PhD, sociologists might think in terms of a population "at risk," out of which persons, over time, move into the status of organic farmer or minority-status PhD. This style of thought applies quite naturally to discrete variables for units of analysis beyond the individual: Among a population of social movements, in the space of movement-time instead of person-time, some movements experience the transition to "defunct." In fact, for most binary variables, sociologists could, as epidemiologists do, recognize that binary variables involve transitions in time from one state to another. This perspective, of course, fundamentally underlies the specific methods of event history, among other techniques.

An underlying focus on the occurrence of events leads many epidemiologists to regard as axiomatic that *rates of occurrence* (hazard rates, in the language of event history) are the fundamental summary measure for binary dependent variables (Rothman 1986). ("Rate" here has its strict sense of events per unit of person-time at risk. See Elandt-Johnson 1975.) Consequently, relationships between independent and dependent variables are conceived in terms of how event rates vary across the independent variable, termed an "exposure" in epidemiology. The epidemiologist thinks of the effect of cigarette-smoking on lung cancer in terms of how rates of cancer occurrence vary across levels of smoking exposure, while sociologists might similarly think of how rates of becoming an organic farmer vary across region or economic circumstances.

In the simplest case, that of a binary independent variable and its effect on a binary dependent variable, effects can be summarized as either a "rate difference" or a "rate ratio." For examining how age-heterogeneous marriage affects the occurrence of the event "a couple divorces," the rate difference would be the rate of divorce among age-heterogeneous couples (perhaps measured as divorces per 1000 age-heterogeneous couples per year), minus the corresponding rate among age-homogeneous couples. The rate ratio (also known as "relative rate") would simply be one rate divided by the other.

The Underlying Logic of Detecting Effects in Case-Control Studies

Case-control designs can produce valid measures of relative effect, such as rate ratios, a fact known to epidemiologists for some 45 years (Cornfield 1951; Corn-

field and Haenszel 1960). Although I do not wish to concentrate here on the properties of statistical estimators, an elementary examination of this topic aids in understanding the logic that underlies case-control studies. Cornfield (1951) used a simple algebraic argument to show that the ordinary odds ratio for a 2 X 2 table obtained via a case-control sample provides a close approximation to the "relative risk," *if one category of the dependent variable is rare,* which is precisely the situation that motivates use of a case-control study. (The "relative risk" is the ratio of conditional percentages in a 2 X 2 table, analogous to the more sociologically familiar difference of conditional proportions.[3]) Further, it is well known that the relative risk approximates the rate ratio under the rare event assumption (e.g., Rothman, 1986). Because an odds ratio remains unchanged if a set of row or column frequencies are multiplied by a constant, a case control study would have the same expected odds ratio, no matter how much the rare category of the dependent variable was oversampled. Estimates of *relative* effect, though *not* measures of *absolute* effect (i.e., the rate difference), are thus possible without knowledge of the population marginals on the dependent variable, which would otherwise appear to be necessary to adjust for stratification. In many case-control studies, knowledge of the marginal distribution on the dependent variable is virtually unobtainable, since the size of the population is unknown.

The contemporary line of thinking about case-control studies has shown that Cornfield's early work, while correct so far as it went, did not fully appreciate the logical foundation of case-control studies and was overly restrictive in requiring the assumption of a rare event. In the 1970s and 1980s, an extensive literature on the logic of case-control studies emerged, the work of Miettinen (1976, 1981, 1982, 1985) being a chief foundation for modern epidemiologic thinking in this area. Deservedly known for its conceptual subtlety, Miettinen's work is nevertheless mathematically simple and provides a relatively easy way to understand the logic and correct practice of case control studies. The portion of Miettinen's methodological work of interest here has stressed issues of logic and proper sample selection, rather than properties of statistical estimators. Thus, his work offers a useful and accessible foundation for the sociologist interested in conducting case-control studies.

To present Miettinen's perspective, let us return to the hypothetical case-control study of factors affecting divorce, with the couple as the unit of analysis. Because of the complexity introduced by repeatable events, well-known both to epidemiologists and to sociological practitioners of event history (Yamaguchi 1991), let us assume a restriction in focus to the *first* divorce of a given couple (leaving out those couples who divorce, remarry one another, and then divorce again). Assume a steady-state, dynamic population. With these assumptions, consider first how the determinants of divorce might be examined using a conventional *longitudinal* study aimed at a comparison of rates. Data would be collected on all members of a defined population at risk, say all married couples who resided in a defined geographic area at some point during a time period of length T. (The following draws from the presentation of Rothman, 1986, as well as

Miettinen's work). Assume that the causal factor of interest is age heterogeneity, defined as couples differing by more than 10 years of age. Let N_D be the number of couples in the study population whose ages differ, and let N_S be the number of couples in which partners' ages are similar. These couples are followed over the period T, and occurrences of divorce are noted. Suppose that D of the N_D age-different couples and S of the N_S age-similar couples divorced during this time period. Such a study would give estimates of the divorce rates (divorces/amount of couple-time observed) in the two groups. The rate among the age-different couples would be $R_D = D/(TN_D)$ and that among the age-similar couples would be $R_S = S/(TN_S)$. The rate difference $(R_D - R_S)$ or rate ratio (R_D/R_S) would offer measures of, respectively, the absolute and relative effect of age difference on the rate of divorce.

Nothing in the preceding should seem surprising or unusual. It is simply one way the results of this panel study might be summarized, given a fundamental focus on rates. Clearly, however, the preceding study would require a lengthy time period or a large panel, given the sparseness of divorce in the space of couple-time. Now, however, let us see how essentially the same results would be obtained, at a fraction of the effort, from a *case-control* study of the same situation. To do this, data on *cases* of divorce are collected retrospectively from the end of the time period. A suitable combing of public records would produce an enumeration of couples who divorced during the period T. By assumption, there is a total of $D + S$ such couples. For comparison, a group of controls is obtained with a random sample of fraction k from the population of all married couples who resided in the community at some point during the period T, whether or not they became divorced during that time. Among the controls, let assume that n_D and n_S couples in the sample were age-different and age-same, respectively. Now consider the data in hand and its relevance to estimating the rate ratio of interest. The enumeration of cases gives D and S, the numerators of the rates. Although denominator information is not known exactly, the controls are a sample of fraction k from the population that yielded the denominators in the longitudinal study. Consequently, the sample estimators of the total number of age-different and age-similar couples in the population would be n_D/k and n_S/k. The corresponding estimated amounts of person-time in the population would be Tn_D/k and Tn_S/k. If k is known, the absolute divorce rates in the two groups can be estimated, as can the rate difference, with the only change from the previously described longitudinal study being that these sample estimates would have to substitute for the population values.

Often, however, knowing the sampling fraction k will be difficult, but it is unnecessary, presuming the rate ratio is satisfactory as a measure of effect. If the rate ratio is estimated as $[D/(Tn_D/k)]/[S/(Tn_S/k)]$, k will cancel out. This rate ratio expression is the same as in the hypothetical longitudinal study of the entire population of interest, except that the sample estimates n_S and n_D substitute for the population values N_S and S_D.

TABLE 1
Hypothetical Data:
Occurence by Age Homogeneity

	Couples' Ages	
	Different	*Similar*
Divorce Cases	D	S
	(a)	(b)
Controls	n_D	n_S
	(c)	(d)

Even more simply, because T also cancels from the estimated rate ratio, if we simply construct a 2 X 2 table, with *Divorce* (Case vs. Control) as the rows, and *Age* (Different vs. Similar) as the columns, the cell entries will be as shown in Table 1, so that the ordinary odds ratio for the sample data $(D/n_D)/(S/n_S) = (ad/bc)$ will also yield the same estimate for the rate ratio. An important and easily overlooked point here is that the controls are *not* a sample from the nondivorced population. To estimate the rate ratio, they must be a sample from the population that generated the cases (couples ever-married and ever resident during period T), that is, the population of couples that would form the denominator of the rates if the longitudinal panel study had been conducted.

This depiction of the essential identity of the results from the case-control study and the longitudinal study nowhere entailed assuming that the event of interest is rare. Thus, Miettinen (1976) showed that an estimate of the rate ratio can efficiently be obtained without the rare event assumption provided that controls are chosen to represent the case-generating population. I use the unusual term "case-generating population" intentionally. This is not identical to "the population of persons who did not experience the event." In fact, some of the controls may have experienced the event but that is irrelevant. This may seem counter-intuitive, but the same phenomenon appears in any ordinary computation of a rate: The denominator of a birth rate, for example, is all persons who could have experienced a birth during the time period in question, not all persons who did *not* experience a birth. If a case-control study is to yield an estimate of the rate ratio without the rare event assumption, the controls must be drawn from the entire population at risk, regardless of whether they experienced the event. Of course, for rare events, the probability of obtaining a control who experienced the event is so low that this distinction makes little difference in practice, but the principle is important.

The irrelevance of the rare event assumption was not fully appreciated until Miettinen's work in the 1970s but has become sufficiently recognized that a recent article specifically treated the use of case-control designs for nonrare events (Rodgrigues and Kirkwood 1990). For the sociologist, this point is of particular importance, since many of our rare events may be less rare than exotic diseases. Most importantly, Miettinen demonstrated that there is no essential logical differ-

ence between case-control and longitudinal (panel) designs, no dubious reasoning from effect to cause. Both kinds of studies estimate a rate ratio or rate difference. Numerators of rates are the same in both studies, but in a case-control study, the denominator (person-time, couple-time, organization-time, etc.) is sampled, rather than known by enumeration of the experience of a panel. This is the essential origin of the economic efficiency of a case-control study—that denominator information is sampled.[4] For relatively more common events, a case-control design may not so markedly improve efficiency, since a sufficient number of events might accumulate in a panel study of smaller size or shorter duration. However, recognizing the irrelevance of the rare event assumption is key to appreciating the modern epidemiologic understanding of case-control studies.

The preceding argument has treated only a simple example, using a binary outcome variable and a single independent variable. However, for more complex substantive models and analyses (e.g., logistic regression), the same logical structure holds: Cases provide information from which numerator information derives, while controls similarly provide the basis for the denominators of relative rate (rate ratio) estimates.

SAMPLING CONTROLS IN CASE-CONTROL DESIGNS

The preceding was intended only as an appreciation of the logic of case-control designs, not as definitive instructions concerning case-control sampling. Appropriate sampling of controls has subtleties glossed over in the divorce example, and I turn to those now.

Incidence Density Sampling

Prior to deeper discussion of sampling, an understanding of the epidemiologic distinction between *incident* and *prevalent* cases is needed. Incident cases are those that develop during some specified time interval, that is, newly occurring transitions into the event status, regardless of whether the case remains so. Prevalent cases are those existing at a given point in time, regardless of when they originated. Taking a sociological example, incident religious cult members newly joined the cult within some defined period of time, while prevalent members are those currently in the cult, regardless of when they joined. In general, using incident cases is preferable (Rothman 1986). While a case-control study may use prevalent cases, even without the rare event assumption, one must assume that the exogenous variables of interest are statistically independent of duration in the event status (Miettinen 1976). Otherwise, the effect of a variable on duration in the status cannot be separated easily from its effect on incidence (Rothman 1986), which biases odds ratio and hence rate ratio estimates. In Eitzen's (1970) study considered above, the "strong Wallace supporters" were prevalent cases, so the associations he found represented a mixture of the effects of status inconsistency on becoming a Wallace supporter and on sustaining that support over time. The

effects could be separated on the assumption that duration of support is unrelated to status inconsistency, an assumption sometimes valid in analogous epidemiologic studies, but such an assumption seems untenable for Eitzen's study. Thus, when possible, incident case series are clearly preferable.

However, when an incident case series is used, sampling of controls becomes more complex. Most simply, as described in the divorce example above, one may sample controls "inclusively" from all persons who were ever in the population during the time period of interest, regardless of when and whether they experienced the event (Rodrigues and Kirkwood 1990). With this kind of sampling of controls, the odds ratio will estimate the rate ratio only if the event is rare; otherwise, it will estimate the relative risk. Or, one may sample cases "concurrently," randomly selecting one or more controls for each case from the population still at risk at the time the case experienced the event (Rodrigues and Kirkwood 1990). Concurrent sampling with an appropriate "matched" analysis (see below) will yield an odds ratio that estimates the rate ratio (relative hazard rate), regardless of whether the event is rare (Rodgrigues and Kirkwood 1990).[5] So, for the example of joining a religious cult, a control might be chosen for each case by random digit dialing, but each potential control would need to be screened to verify that they were resident in the community and not yet cult members when the corresponding case joined the cult. Another approach, described by Walker (1994), suggests a two-step sampling process: (a) A date is selected at random from the time period in which cases emerged; and (b) Persons are randomly selected from the population, and accepted as controls if they were in the population as of that date. These steps are repeated until the desired number of controls is achieved. With this method, when independent variable data are collected from cases with reference to the time at which they became cases, the corresponding data are collected from controls relative to their random dates: If incident religious cult members are asked "In the year prior to when you joined ...," controls would be asked "In the year prior to [random date], ..." Whether concurrent sampling or the method described by Walker is used, the sample of controls is not simply a sample of persons, but a sample of person-time, since it can be shown that the probability of being selected as a control is "proportional to the duration of time that the individual spends in the source population at risk" (Walker 1994:76). Because such methods sample person-time at risk and thereby permit estimates of incidence rate ratios, they are referred to as "incidence density" sampling (Miettinen 1976).

Notwithstanding the preceding, an investigator who is satisfied with a valid odds ratio that does not necessarily approximate the rate ratio may simply sample controls from persons who never experienced the event during the time frame of interest (Rodrigues and Kirkwood 1990). This sampling design yields a more powerful estimate of effect (Rodrigues and Kirkwood 1990), but since persons rather than person-time at risk has been sampled, the connection to rate ratios (relative hazard rates) and event history is lost, unless the event is rare.

As in any sample-based study, increasing sample size will increase precision, that is, reduce sampling error. Assuming an enumeration of all cases, only the

sample size of controls can be increased, so case-control studies usually involve at least as many controls as cases. Increasing precision with a larger sample of controls is subject to diminishing returns. As a rule of thumb, relative precision reaches about 90 percent of its theoretical maximum with a control to case ratio of 4-5:1 (Lasky and Stolley 1993). A larger ratio might nevertheless be desirable for the investigation of sparsely-distributed independent variables, or if a multivariate analysis with many cells is to be performed.

In a further extension from the procedures described above, it is worth noting that a complete enumeration of cases may be unnecessary or undesirable. In a case-control study of divorce in a large city, the population of available cases would be sufficiently large that a one year enumeration of cases would exceed what is needed. One might restrict the study to divorces occurring during a shorter period, but this might be undesirable, should factors leading to divorce vary across times of the year. In that situation, a random sample of divorces across the year might be better. Just as the sampling fraction for controls cancels out of the odds ratio, leaving a rate ratio estimate, so too will the sampling fraction for cases.

Using Nonrandom Controls to Approximate the Case-Generating Population

Another modification of sampling procedures often used by epidemiologists involves a judgment sample of controls, and this tactic may have applicability in sociological case-control studies. In a study of incident heart attack cases consecutively arriving for treatment at Hospital X (in a locale with several other hospitals), an epidemiologist might select as controls a series of patients who came to Hospital X with some other disease during the same time period, rather than sampling the general population of the local geographic area. Such a choice is not purely one of convenience, but instead rests on judgments of representativeness. Because different hospitals draw patients with different social and economic characteristics, a sample from the general population will poorly represent the population that generated the cases at Hospital X. Thus, the epidemiologist's rationale for using this "judgment sample" of controls would be that persons arriving at Hospital X with another disease would better approximate the case-generating population than would a sample from the "general" population. In situations in which it is impossible to obtain any sort of listing of the case-generating population, or to target them through a procedure such as random-digit dialing, a nonrandom judgment sample of controls may not only be more convenient, but may be more representative as well.

To better understand this sort of sampling, and to recognize its potential risks and benefits, notice that the only function of controls is to provide information on the distribution of the independent variable within the case-generating population. For example, suppose that one wishes to examine age as a causal factor in the hypothetical study of heart attack patients at Hospital X. In principle, then, any sample of controls that provides accurate information on the distribution of the independent variable in the case-generating population will be satisfactory.

However, consider how a bias might occur: If the epidemiologist chose as controls a series of patients who came to the hospital for treatment of athletic injuries, a bias will ensue, since persons with athletic injuries will likely have an age distribution quite different from the case-generating population for heart attack patients. The general condition for *valid* use of a nonrandom control sample is that it must be selected independently of the causal variable (s) relevant to the event of interest (Rothman 1986). Thus, in the current illustration, controls must be selected from among patients who arrive with a disease event *known* to be independent of age. This may be a difficult condition for social science studies to meet in analogous situations, since it depends on a background of theoretical and empirical knowledge that permits prior awareness of when such dependence is likely to exist.

Nevertheless, nonrandom control samples might have a place in sociology. For example, in Eitzen's (1970) study described previously, the event of interest, technically speaking, was "being noticed as a passing car with a Wallace bumper sticker." Thus, rather than using a sample from local auto registration rolls (the "general population"), Eitzen might have approximated the case-generating population by selecting the controls from the owners of other passing cars. Because the causal factor of interest in Eitzen's study was status consistency, the condition for valid selection of controls would be that their probability of selection be independent of (unrelated to) their status consistency. While it seems reasonable that this condition could be met, there is certainly room for error. Suppose that passing-car controls were selected (for convenience) at particular times of day or street locations in the community. This would be analogous to selecting persons with some particular other disease as controls, and the potential source of bias would be that the status consistency of drivers might be related to when or where they typically drove their cars. If this kind of dependency did exist, the probability of car owners being selected as controls would be related to their status consistency. Thus, investigators contemplating nonrandom control samples need to anticipate what independent variables are theoretically relevant, and choose controls so that their selection is independent of any of the independent variables of interest.

Multiple Independent Variables and the Selection of Controls

So far the discussion has proceeded as though case-control studies can only consider the effect of a single variable on the event of interest. To the contrary, carefully executed case-control studies are valid and efficient for exploring multiple independent variables. In a longitudinal study exploring the effect of several different independent variables on the hazard rate for a rare event, a panel large and diverse enough to exhibit a reasonable distribution of all the independent variables among both event and nonevent subjects would be needed. Follow up costs for such a panel will likely be quite high. What if the initial AIDS investigators had mounted a longitudinal exploratory study to discover the cause and mode of transmission of AIDS? In a case-control study, by contrast, the much

larger sample of persons who experienced the rare event allows for statistically precise estimates for a variety of independent variables. With such flexibility and apparent ease comes risk, however, as the validity of effect estimates depends on the controls having been sampled independently of each of the causal variables of interest. A randomly sampled control series from a well defined population at risk fulfills this condition. But if a nonrandom control group is used, the chance of bias increases, since it may accurately represent the distribution for one independent variable, but not others (For a substantive example of this kind of problem in epidemiology, see MacMahon, Yen, Trichopoulos, Warren, and Nardi 1981; Feinstein, Horowitz, Spitzer, and Battista 1981).

Matched Samples in Case-Control Studies

Controls matched to cases on one or more variables are often used in case-control studies. Even some of the few sociological case-control studies have used matching. For example, Kruttschnitt (1989) conducted an innovative study comparing convicted violent criminal offenders (cases), to controls individually matched on race, age, and neighborhood of residence in adolescence.

The function and proper use of matched control samples in case-control studies does not follow the ordinary intuitions that would come from experience with matching in experimental or panel studies, in which matching eliminates confounding or spurious effects associated with the matching variable(s). In case-control studies, matching actually can create confounding that otherwise would be absent and generally biases (attenuates) effect estimates towards the null, regardless of the direction of the true effect (Rothman 1986; Breslow and Day 1980). This bias can and must be eliminated in case-control studies by controlling for the matched variable during the stage of data analysis, and will be discussed below. At this point, it is sufficient to note that the use of matched controls without appropriate statistical adjustment has appeared in at least one sociological case-control study (Kruttschnitt 1989), and probably reduced the strength of relationships found in the data analysis.

Matching does have a useful role in case-control studies by enhancing the precision with which the effect of a confounding variable can be controlled in situations in which the population of cases and controls differ substantially in their distributions on the confounder variable. Imagine a case-control study aimed at studying the effect of having a law degree on the attainment of political office, with the analyst planning to control for gender in the analysis. Let us assume that females will be of low frequency among the cases of new officeholders. If an investigator failed to match on gender in selecting controls, roughly half of the controls will be female, a proportion greatly exceeding what will prevail among the cases. When gender is used a control variable, the females' subtable showing New Officeholder X Law Degree will have a considerable excess of controls relative to new officeholders, while the comparable subtable for males will have few controls, relative to cases. Here, ordinary intuitions about statistical precision are correct: Having relatively comparable numbers of cases and controls

in the partial tables would be preferable. More nearly optimal allocation of subjects for the purposes of statistical control would be achieved if cases and controls were matched on gender during selection. This ensures balanced numbers of cases and controls in both the male and female subtables.

Thus, while selecting matching controls does not by itself eliminate the effect of nuisance variables, it can assist in obtaining data with a frequency distribution that facilitates precise statistical control. But matching will not always increase statistical efficiency (Thomas and Greenland 1983), and matching is typically expensive, so its use should be carefully considered in advance.

DATA ANALYSIS FOR CASE-CONTROL STUDIES

Up to this point, discussion of data analysis has been limited to that necessary for explanation of sampling techniques and for understanding the logic of case-control designs. In this section, I summarize some principles and techniques of data analysis, and point the reader toward sources that offer further detail.

As a general principle, case-control data must be analyzed using measures of effect or association that remain unchanged if the marginals on the dependent variable are multiplied by a constant. This necessity follows from the fact that case-control studies involve stratification and oversampling on the dependent variable, without knowledge of the population distribution to permit adjustment. This feature rules out a conventional tabular analysis of differences of conditional proportions or percentages, because any multiplicative alteration of the dependent variable's marginals will change the difference in proportions. For example, consider the hypothetical divorce data in Table I. The difference in conditional proportions of divorced couples would be $[a/(a+c)] - [b/(b+d)]$. However, suppose the investigator had drawn a control sample that was twice as large. This would double the size of d and c, so that the measure of effect, in the absence of weighting to adjust for sampling ratios in the numerator and denominator, would be a function of the size of the control sample. This is clearly undesirable.

By contrast, measures of effect that do not vary with respect to the size of the marginals on the dependent variable will work well even in the absence of weighting. For example, consider Goodman and Kruskal's Gamma, which for Table 1 will be equal to $(ad - bc)/(ad + bc)$. This expression stays the same if c and d are multiplied by any constant, which simulates what would happen when c and d cannot be weighted to reflect disproportionate sampling fractions. Besides Gamma and similar measures, another and more complex alternative would be to analyze the data using variety of techniques developed in the econometric literature on "choice-based" or "response-based" sampling (Coslett 1981; Manski 1995; Manski and Lerman 1977).

Rather than using PRE measures of association or the econometric techniques, I would recommend that sociologists interested in analyzing case-control data would do well to use odds-based techniques, as have been emphasized here. There are several reasons for this: The odds ratio is a convenient measure that is

invariant with respect to multiplicative alterations of the marginal distribution on the dependent variable. Sociologists familiar with loglinear techniques, including logistic regression, can analyze case-control data without having to learn any fundamentally new techniques. Finally, with appropriate sampling as described above, odds ratios obtained from a simple tabular analysis, from a loglinear analysis, or from a logistic regression will all yield valid estimates of the rate ratio (relative hazard rate), assuming that the rate ratio is constant over the time period under study.[6] There are a few caveats: In logit or logistic regression models, the intercept coefficient will not be correct, so that probabilities or absolute rates or risks cannot be inferred, but exponentiating the coefficients will yield the odds ratios associated with the independent variables, which, as already seen, will give estimates of rate ratios.[7] Thus, for many case-control studies, the quantitative sociologist comfortable with modern odds-based techniques not need learn much in the way of new statistical tools to analyze case-control studies.

An exception to the preceding claim of familiarity, though, comes when matched controls are used. As already indicated, using matched controls requires a matched analysis, the more advanced forms of which are likely to be unfamiliar to sociologists. Even when matching in the conventional sense is absent, it may enter a case-control study through concurrent sampling of controls (Rodrigues and Kirkwood 1990), which makes controls time-matched to their respective cases. A matched analysis is necessary for this situation, as well as for the more familiar matching of cases and controls on gender, ethnicity, and so forth. If the matching variable(s) only have a few categories, simply using them as controls in a tabular analysis, or entering gender, ethnicity, etc. as an independent variables in a loglinear or logistic regression model will be sufficient to provide an odds ratio estimate that controls for the matching procedure. This is no different that what might be done with any covariable in a multivariate analysis. While many sociologists would use loglinear methods, tabular techniques for obtaining estimates of an odds ratio net of the effect of control variables are not complex, are discussed in standard epidemiology and categorical data analysis textbooks (e.g., Schlesselman 1982; Rothman 1986; Agresti 1990), and are available in some standard statistical packages.[8]

However, when sampling involves individual matching of cases and controls on multiple variables or a variable with numerous values, methods less familiar to sociologists are required. In this situation, the matching variable will have as many values as there are cases. In the simplest situation of one control matched to each case, when the analyst controls for pair matching, each subtable in a tabular analysis will contain only two paired individuals, one case and one control. If the independent variable is binary, the ratio of the number of subtables with discordant matched pairs (case and control differ on the independent variable) to the number of concordant pairs will estimate the odds ratio and hence the rate ratio. Any detailed discussion of this situation, as well as the more complicated one of multiple matched controls per case is beyond the scope of the current paper. However, the interested reader can find helpful material on tabular matched anal-

yses in various standard epidemiologic sources, such as Breslow and Day (1980) and Rothman (1986). When individual matching is used, but the independent variable of interest is continuous, conditional maximum likelihood logistic regression analysis is necessary (Hosmer and Lemeshow 1989). For matched pairs, this analysis can be performed by applying certain coding "tricks" to conventional logistic regression programs, as described by Hosmer and Lemeshow (1989), but the analysis is easier with a specialized program (Campos-Filho and Franco 1989), which also will work if more than one control is matched to each case. A matched pair analysis of this kind would have would have been appropriate for Kruttschnitt's (1989) study described previously. In summary, then, case-control studies with individual matching will require that the investigator gain familiarity with new if not particularly complex statistical techniques.

AN EMPIRICAL ILLUSTRATION: CHANGE OF CHIEF ADMINISTRATORS AT EDUCATIONAL INSTITUTIONS

To illustrate some of the principles of design and analysis detailed above, I conclude with an empirical example of a case-control study. The event of interest is "acquiring a new chief administrative officer" (president, chancellor, etc.; henceforth, "CAO") among post-secondary educational institutions. The case-generating population included all universities, colleges, community colleges, technical schools and so forth extant in the United States in 1992-1993 (N = about 3700). Each yearly volume of the *Peterson's Register of Higher Education* (Peterson's Guides 1993) offers a brief sketch of all extant institutions, including such characteristics as enrollment, degrees offered, and tuition. Its relevance here is that it conveniently lists all institutions that experienced a change of CAO during the preceding year.

During 1992-1993, 254 institutions acquired a new CAO, and constituted the cases for this study. Because the case-defining event occurred (or at least was reported) in discrete time, on a yearly basis, none of the time-related complexities of sampling controls were at issue. I selected a simple random sample of controls (N = 254), without matching, from the population at risk, which was all institutions extant in 1992. For each case and control, the following variables were obtained: Enrollment; Institutional Status (Two-Year, Four-Year, Comprehensive, University, Graduate Only); Type of Administrative Control (State, Independent Nonreligious, Local Government, Independent-Religious, Proprietary), and Tuition (minimum yearly tuition for undergraduates).

Many of these variables had little if any effect on the rate ratio for change of chief administrative officer. For illustration, I have focused on variables that did have an association with change in CAO. Table 2 shows simple tabular analyses, using the odds ratio (OR) as the measure of effect. On the rationale that different types of administrative control (State Government, Proprietary, etc.) provide differing career opportunities for CAOs, and create different organizational cultures, I examined the effect of Type of Administrative Control on the rate ratio

TABLE 2
Change of CAO by Type of Administrative Authority

Type of Control	*Changed CAO*	*Controls*	*OR**	*95% C.I.*[†]
Local Government				
Yes	27	40	0.64	0.38, 1.07
No	227	214	$p = 0.088$[‡]	
State Government				
Yes	72	80	0.86	0.59, 1.26
No	182	174	$p = 0.44$	
Proprietary				
Yes	31	18	1.82	0.99, 3.35
No	223	236	$p = 0.051$	

Notes: *Odds ratio.
[†]95% confidence interval for the odds ratio.
[‡]*p*-values here and below are two-sided.

for change in CAO. The results in Table 2 indicate that Local Government participation in institutional administration, reduces the odds ($OR = 0.64$) of experiencing a change in CAO.[9] Because the sample of controls was drawn from the population at risk, the odds ratio estimates the rate ratio, so this result can be interpreted as a rate ratio, indicating that the estimated rate of change of CAO is only 64 percent as high among institutions with Local Control as it was among other institutions. Similarly, State Government Control slightly tended to reduce the rate of CAO turnover ($OR = 0.86$) from what it was among other institutions. Proprietary institutions had an increased rate of CAO change, relative to other kinds of institutions ($OR = 1.82$).

Although these tabular analyses show a simple and clear method to obtain odds ratios and therefore rate ratio estimates, logit analyses produce essentially similar results. To illustrate a logit analysis, I analyzed a three variable model, using the SPSS-Windows loglinear procedure (Norusis 1993b). Because institutions with Proprietary Control are much more likely than others to have Two-Year Status (80% vs. 34% in the current sample), the effect of Two-Year Status may confound or suppress the effect of Proprietary Control. Table 3 contains the results relevant to this supposition. The first logit model includes only the effect of Proprietary Control. With appropriate choice of contrast coding, the exponentiated logit coefficient yields an odds ratio (1.82) and confidence interval identical (to two decimal places) to that obtained in the preceding tabular analysis. The second logit model adds in the main effect of Two-Year Status. This model indicates that while Two-Year Status did not itself have a statistically reliable effect (the confidence interval on the logit coefficient includes 0), Two-Year Status did suppress the effect of Proprietary Control, which in this model shows a slightly larger odds ratio (relative rate) of 2.01. I also ran the saturated logit model, allowing for an interactive effect of Two-Year Status and Proprietary Control in relation to CAO change, but this effect was insubstantial and did not change the estimated

TABLE 3
Logit Analysis of Propritary Control, Two-Year Status, and Change of CAO

Variable	*Logit Coeff.*	*Odds Ratio**	*95% C.I.*[†]
Model 1: {Proprietary Control, Change of CAO}			
Proprietary Control	0.60	1.82	0.99, 3.35
$p = 0.51$, Likelihood Ratio χ^2, $df = 2$			
Model 2: {Proprietary Control, Two-Year Status, Change of CAO}			
Proprietary Control	0.70	2.01	1.07, 3.81
Two-Year Status	-0.22	0.80	0.55, 1.17
$p = 0.829$, Likelihood Ratio χ^2, $df = 1$			

Notes: *Odds ratio = exp (Logit Coeff.).
[†]95% confidence interval for the odds ratio.

effect of Proprietary Control. Thus, Proprietary Control has a positive effect on the relative rate of CAO change, and the partial effect is actually higher than the zero-order effect, having been suppressed the strong association of Proprietary Control with Two-Year Status, which has a weak negative effect on the relative rate of CAO change.

A final set of analyses, in Table 4, illustrates the use of logistic regression to accommodate continuous as well as categorical covariates. Enrollment, scaled to units of 10,000, was introduced as an independent variable, on the assumption that institutions with larger enrollments offer more prestigious and lucrative positions for CAOs, and should therefore be more successful in retaining CAOs. Model 1 in Table 4 indicates that higher Enrollment reduces the relative rate of CAO change by a factor of 0.76 for every one unit (10,000 student) increase in enrollment. The estimated rate ratio for Proprietary Control remained approximately the same ($OR = 1.9$) as shown in Tables 2 and 3. Model 2 in Table 4 adds

TABLE 4
Logistic Regression Models for the Effect Of Enrollment, Proprietary Control and Two-Year Status on Change of CAO

Variable	*Logit Coeff.*	*Odds Ratio**	*95% C.I.*[†]
Model 1.			
Proprietary Control	0.64	1.89	0.97, 3.71
Enrollment[‡]	-.27	0.76	0.56, 1.02
$p = 0.017$, Likelihood Ratio $\chi^2 = 8.16$[§], $df = 2$			
Model 2.			
Enrollment	-0.27	0.76	0.57, 1.02
Proprietary Control	0.72	2.05	1.02, 4.12
Two-Year Status	-0.19	0.83	0.56, 1.21
$p = 0.028$, Likelihood Ratio $\chi^2 = 9.13$, $df = 3$			

Notes: *Odds ratio [exp (Coeff.)] associated with a one unit change in each regressor.
[†]95% confidence interval for the odds ratio.
[‡]Enrollment scaled to units of 10,000.
[§]χ^2 values obtained from comparison against an intercept-only model.

Two-Year Status to the variables in Model 1. Although the improvement in goodness of fit over Model 1 is moderate (difference of Model $X^2 = 0.96$, $df = 1$, $p = 0.33$), the partial effect of Proprietary Control again increased to an estimated rate ratio of 2.05, while the rate ratio associated with Enrollment remained unchanged. The estimated partial effect of Two-Year Status, though quite imprecise as indicated by its wide confidence interval, is negative and approximately the same as was found in the previous analyses in Table 3.

DISCUSSION AND CONCLUSIONS

The preceding empirical example demonstrates a situation in which case-control methods offered a considerable economy of effort. Although assembling data on *all* of the 3700 or so institutions of higher education extant in the United States in 1992-1993 would have been possible, case-control sampling reduced the sample size to 508. The more time-consuming option of collecting data on the entire population would only have eliminated sampling error in the denominator, that is, in the case-generating population. Alternatively, had I drawn a simple random sample of 508 from the entire case-generating population, with no stratification on the dependent variable, the expected number of events ("cases") in the sample would have been far too small for a precise analysis: Assuming a total population size of 3700, the proportion of institutions that experienced a change of CAO would be about 7 percent (254 cases/3700). Applied to a simple random sample of 508, the expected number of CAO cases per year would have been 35. The use of a case-control design focused data collection effort where it offered the largest return in precision, on the events (numerator information).

This illustrates the central argument in favor of case-control studies. For studying low-frequency events, case-control design offers economy of effort without loss of validity, without illogical reasoning from effect to cause, and without resort to highly specialized or unusual methods of analysis. Nor is it necessary to develop from scratch a new set of methodological principles, as the ground has been well-covered by epidemiologists for many decades.

Of course, not all rare events are equally amenable to study via case-control design. Its efficient use requires that it be possible to obtain an enumeration or sample of events. Consequently, economies of design will be greatest for events subject to governmental record-keeping or events whose occurrence otherwise leaves a public trace. Many such events interest sociologists: Crimes, marriages, corporate mergers, changes of political regime, and attainment of political office illustrate a few. In other situations, such as the religious cult example discussed above, enumeration of persons or other units exemplifying the event would be possible by simply contacting the relevant organization, even though no official public record would exist. Thus, while case-control design cannot solve all the difficulties of studying rare events, it is applicable to many sociological topics.

By offering an introduction to the potential and logic of case-control designs, and a basic description of how properly to conduct them, I hope this essay will

further their use in sociology. Response-based or case-control designs need not be the special practice of only the most technically proficient sociologists, nor need they be used in a semiconscious way, as occurred in the past. The rise of case-control methodology revolutionized epidemiologic research over a few decades. Because of the difference in disciplines, no such revolution will be forthcoming in sociology. Nevertheless, many substantively valuable sociological studies could emerge if case-control design became part of the standard methodological apparatus of the ordinary quantitative sociologist.

Acknowledgements: I thank Stephan Lanes for introducing me to case-control studies, for his collaboration on an earlier related paper, and for numerous conceptual clarifications over the last several years. Kashinath Patil gave prompt and focused advice concerning the treatment of matched designs. Prabha Unnithan offered several helpful comments on a draft of the manuscript. Dennis Roncek helped me work out a general strategy for the presentation of the ideas here.

NOTES

1. By person-time, I mean the composite of some number of persons and the times over which they are observed (Walker 1994). Although this concept clearly applies to units of analysis other than the individual, I often use "person-time" here for convenience and without any intended loss of generality. This avoids clumsy locutions like "unit of analysis-time."
2. Finding sociological examples of such designs is difficult, given that references to methods may not show up through a keyword search of titles and abstracts. Therefore, to look for previous sociological uses of these designs, I canvassed the *Social Science Citation Index* from 1985 to the present. On the rationale that articles using or discussing such designs would be revealed by citations to the classic methodological articles, I searched for citations to the econometric literature that has appeared in sociological venues (Manski 1981; Xie and Manski 1989), to the founding econometric literature that has appeared elsewhere (Manski and Lerman 1977; Manski and McFadden 1981; McFadden 1973), and to the two articles on case-control designs in criminology (Goodman, Mercy, Layde, and Thacker 1988; Loftin and McDowell 1988). I also used "case-control,", "response-based," "choice-based," "retrospective", and "discrete choice" as keywords in title searches. This search revealed one response-based substantive article by a sociologist since 1985 (Arnold and Hagan 1992), and I have encountered three other response-based investigations by chance (Eitzen 1970; Gortmaker 1979; Kruttschnitt 1989). All of these will be discussed later in the paper. There were many internal citations to the econometric literature that has appeared in sociology (e.g., Xie and Manski 1989, citing Manski 1981), but these were themselves purely methodological as opposed to substantive articles, dealing with difficult technical points. Besides casual observation of sociologists' practice, this search of the journal literature is the basis on which I claim that use of and knowledge about case-control designs is rare in sociology.

3. More exactly, in a 2 X 2 table with first row frequencies a, b, and second row frequencies c, d, the relative risk or "risk ratio" is $[a/(a+c)]/[b/(b+d)]$. This statistic is popular in epidemiologic applications because the difference in conditional proportions is unrevealing when event frequencies (a and b) are low. Consider a case in which the relative frequency of an event is 0.001 among males, and 0.01 among females. (Sexual assault victimization might be a relevant example.) The difference in conditional proportions will be less than 0.01, but the relative risk will be 10.0, indicating a strong effect of gender. For more detail, see any standard epidemiologic source, such as Rothman (1986), Schlesselman (1982), or Walker (1982).
4. Practical considerations other than efficiency may make a case-control study preferable. Rodrigues and Kirkwood (1990) point to the ethical advantages of case-control studies over random assignment and longitudinal followup in examining the efficacy of certain common medical treatment regimes (e.g., immunization). Conceivably, there may be parallel situations in social program evaluation.
5. To understand why it makes a difference whether the controls are sampled inclusively or concurrently and whether the event of interest is rare, it is helpful to emphasize that the odds ratio estimates a rate ratio only when the controls represent *person-time* at risk, not just persons at risk. This follows from the fact that rates measure events per unit of person-time, whereas risks (relative frequencies) measure events per person. With inclusive sampling (by contrast to concurrent sampling), the probability of selection for controls in no way reflects how long they were at risk. Instead, the probability of selection is the same for everyone in the population, so that odds ratios will estimate risk ratios. By contrast, with concurrent sampling, the probability of selection as a control will be proportional to time spent at risk (i.e., time spent in the population while still not having experienced the event). Thus, controls in concurrent sampling represent person-time at risk, and the odds ratio will estimate the rate ratio. However, this distinction matters little for rare events. Consider a case-control study of a rare event such as homicide victimization among the U.S. population, carried out over a one-year period. Because so few persons will become victims, essentially all persons in the population are continuously at risk (i.e., have not yet been murdered) for the entire one-year period. Consequently, there will be no practical difference between selecting controls from among people who are still at risk at any given point in time (concurrently), and selecting controls from among all persons in the population (inclusively).
6. This is the same as the assumption of proportional hazards in event history, as noted by Rodrigues and Kirkwood (1990).
7. However, Xie and Manski (1989), have contributed a note of caution concerning the use of logit models in response-based samples. They have shown that if the true specification is not logit, conventional logistic regression gives biased coefficient estimates for large samples. For those situations, they recommend the weighted maximum likelihood estimator of Manski and Lerman (1977).
8. SAS (1989) generally has had more features conducive to analysis of case-control data than has SPSS (Norusis 1993a, 1993b), since SAS has long served a biostatistical as well as social science community. However, differences between the two have considerably narrowed in recent years and will undoubtedly continue to do so in the future. A shareware package with many convenient features for tabular analysis of case-control data is Epi Info (Dean, Dean, Coulombier, Brendel, Smith, Burton, Dicker, Sullivan, Fagan, and Arner 1994), available via the World Wide Web from the U.S. Centers for Disease Control and Prevention.

9. I used Epi Info (Dean et al., 1994) to obtain odds ratios, confidence intervals and *p*-values for these tabular analyses. Similar results can be obtained from SPSS (Norusis 1993a), but Epi Info offers exact tests and confidence intervals, which I have reported in lieu of asymptotic approximations whenever the exact and asymptotic results substantially differed.

REFERENCES

Agresti, Alan. 1990. *Categorical Data Analysis*. New York: John Wiley & Sons.

Allison, Paul D. 1984. *Event History Analysis*. Beverly Hills: Sage Pubns.

Arnold, Bruce L., and John Hagan. 1992. "Careers of Misconduct: The Structure of Prosecuted Professional Deviance among Lawyers." *American Sociological Review* 57:771-780.

Booth, Alan, David Johnson, Lynn K. White, and John N. Edwards. 1991. "Marital Instability over the Life Course: Methodology Report for a Three-Wave Panel Study." Department of Sociology, University of Nebraska, Lincoln, NE.

Breslow, Norman E., and Day, N. E. 1980. *Statistical Methods in Cancer Research. Volume 1. The Analysis of Case Control Studies*. IARC Scientific Publications No. 32. Lyon: International Agency for Research on Cancer.

Bye, Barry. V., Salvatore J. Gallicchio, and Jesse M. Levy. 1987. "Estimation of Discrete Choice Models in Retrospective Samples: Application of the Manski and McFadden Conditional Maximum Likelihood Estimator." *Sociological Methods and Research* 15:467-92.

Campos-Filho, Nelson, and Eduardo L. Franco. 1989. "A Microcomputer Program for Multiple Logistic Regression by Unconditional and Conditional Maximum Likelihood Methods." *American Journal of Epidemiology* 129:439-444.

Cornfield, Jerome. 1951. "A Method of Estimating Comparative Rates from Clinical Data." *Journal of The National Cancer Institute* 11:1269-75.

Cornfield, Jerome, and William Haenszel. 1960. "Some Aspects of Retrospective Studies." *Journal of Chronic Disease* 11:523-34.

Coslett, Stephen R. 1981. "Efficient Estimation of Discrete-Choice Models." Pp. 51-111 in *Structural Analysis of Discrete Data with Econometric Applications*, edited by Charles Manski and Daniel McFadden. Cambridge: MIT Press.

Dean, Andrew. G., Jeffrey A. Dean, Denis Coulombier, Karl A. Brendel, Donald C. Smith, Anthony H. Burton, Richard C. Dicker, Kevin Sullivan, Robert. F. Fagan, and Thomas G. Arner. 1994. *Epi Info, Version 6: A Word Processing, Database, and Statistics Program for Epidemiology on Microcomputers*. Atlanta: Centers for Disease Control and Prevention.

Diekmann, Von Andreas, and Klein, Thomas. 1991. "Bestimmungsgründe des Ehescheidungsrisikos: Eine Empirische Untersuchung mit den Daten des Sozioökonomischen Panels." *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 43:271-290.

Durkheim, Emile. 1951. *Suicide*. Glencoe, IL: Free Press.

Eitzen, D. Stanley. 1970. "Status Inconsistency and Wallace Supporters in a Midwestern City." *Social Forces* 48:493-498.

Elandt-Johnson, Regina C. 1975. "Definitions of Rates: Some Remarks on their Use and Misuse." *American Journal of Epidemiology* 102:267-271.

Feinstein, Alvan R., Ralph I. Horowitz, Walter O. Spitzer, and Renaldo N. Battista. 1981. "Coffee and Pancreatic Cancer. The Problems of Etiologic Science and Epidemiologic Case-Control Research." *Journal of the American Medical Association* 246:957-961.

Fienberg, Stephen E. 1980. *The Analysis of Cross-Classified Data*, 2nd ed. Cambridge: MIT Press.

Glueck, Sheldon, and Eleanor Glueck. 1950. *Unraveling Juvenile Delinquency*. Cambridge: Harvard University Press.

Goodman, Richard A., James A. Mercy, Peter M. Layde, and Stephen B. Thacker. 1988. "Case-Control Studies: Design issues for Criminological Applications." *Journal of Quantitative Criminology* 4:71-83.

Gortmaker, Steven L. 1979. "Poverty and Infant Mortality in the United States." *American Sociological Review* 44:280-297.

———. 1981. "Some Useful Applications of Logistic Models: Reply to an Odd Critique." *American Sociological Review* 46:943-944.

Hirschi, Travis, and Hanan Selvin. 1967. *Principles of Survey Analysis*. Glencoe, IL: Free Press.

Hosmer, David W., Jr., and Stanley Lemeshow. 1989. *Applied Logistic Regression Analysis*. New York: Wiley.

Kruttschnitt, Candace 1989. "A Sociological, Offender-Based, Study of Rape." *Sociological Quarterly* 30:305-310.

Lasky, Tamar, and Paul D. Stolley. 1993. "Selection of Cases and Controls." *Epidemiologic Reviews* 16:6-17.

Loftin, Colin, and David McDowall. 1988. "The Analysis of Case-Control Studies in Criminology." *Journal of Quantitative Criminology* 4:85-98.

MacMahon, Brian, Stella Yen, Dimitrios Trichopoulos, Kenneth Warren, and George Nardi. 1984. "Coffee and Cancer of the Pancreas." *New England Journal of Medicine* 31:430-434.

Manski, Charles F. 1981. "Structural Models for Discrete Choice." Pp. 58-109 in *Sociological Methodology 1981*, edited by Samuel Leinhardt. San Francisco: Jossey-Bass.

Manski, Charles F. 1995. *Identification Problems in the Social Sciences*. Cambridge: Harvard University Press.

Manski, Charles F., and Steven R. Lerman. 1977. "The Estimation of Choice Probabilities from Choice-Based Samples." *Econometrica* 45:1977-1989.

Manski, Charles F., and Daniel McFadden. 1981. "Alternative Estimators for Discrete Choice Analysis." Pp. 2-50 in *Structural Analysis of Discrete Data with Econometric Applications*, edited by Charles Manski and Daniel McFadden. Cambridge: MIT Press.

McFadden, Daniel. 1973. "Conditional Logit Analysis of Qualitative Choice Behavior." Pp. 105-142 in *Frontiers in Econometrics*, edited by Paul Zarembka. New York: Academic Press.

Miettinen, Olli S. 1976. "Estimability and Estimation in Case-Referent Studies." *American Journal of Epidemiology* 103:226-35.

———. 1981. "An Alternative to the Proportionate Mortality Ratio." *American Journal of Epidemiology* 114:114-48.

———. 1982. "Design Options in Epidemiologic Research." *Scandinavian Journal of Worker and Environmental Health* 8 suppl. 1:7-14.

———. 1985. "The 'Case-Control' Study: Valid Selection of Subjects." *Journal of Chronic Disease* 38:543-48.

Norusis, Maria J. 1993a. *SPSS for Windows, Base System User's Guide, Release 6.0.* Chicago: SPSS Inc.

———. 1993b. *SPSS for Windows, Advanced Statistics, Release 6.0.* Chicago: SPSS Inc.

Peterson's Guides. 1993. *Peterson's Register of Higher Education, 1992.* Princeton, NJ: Peterson's Guides.

Rodrigues, Laura, and Betty R. Kirkwood. 1990. "Case-Control Designs in the Study of Common Diseases: Updates on the Demise of the Rare Disease Assumption and the Choice of Sampling Scheme for Controls." *International Journal of Epidemiology* 19:205-213.

Rothman, Kenneth. 1986. *Modern Epidemiology.* Boston/Toronto: Little-Brown.

SAS Institute, Inc. 1989. *SAS/STAT User's Guide. Version 6,* 4th ed. Cary, NC: SAS, Institute, Inc.

Schlesselman, James J. 1982. *Case-Control Studies: Design, Conduct, Analysis.* New York: Oxford University Press.

Swafford, Michael. 1980 ."Three Parametric Techniques for Contingency Table Analysis: A Nontechnical Commentary." *American Sociological Review* 45:664-90.

Swafford, Michael. 1981. "Mortality Rates in a Two-Sample Study." *American Sociological Review* 46:944-946.

Thill, Jean-Claude, and Joel L. Horowitz. 1991. "Estimating a Destination-Choice Model from a Choice-Based Sample with Limited Information." *Geographical Analysis* 23:298-315.

Thomas, Donald C., and Sander Greenland. 1983. "The Relative Efficiencies of Matched and Independent Sample Designs for Case-Control Studies." *Journal of Chronic Disease* 36:685-697.

Tuma, Nancy B., and Hannan, Michael T. 1984. *Social Dynamics.* New York: Academic Press.

U.S. Bureau of Census. 1995. *Statistical Abstract of the United States, 1995.* Washington, DC: Government Printing Office.

Walker, Alexander M. 1991. *Observation and Inference: An Introduction to the Methods of Epidemiology.* Newton Lower Falls, MA: Epidemiology Resources Inc.

Xie, Yu, and Charles F. Manski. 1989. "The Logit Model and Response-Based Samples." *Sociological Methods and Research* 17:283-302.

Yamaguchi, Kazuo. 1991. *Event History Analysis.* Applied Social Research Methods Series, Volume 28. Newbury Park, CA: Sage.