CROSS-SECTIONAL AND LONGITUDINAL DESIGNS

Martin A. Kozloff

A. Cross-sectional

A cross-sectional design (often in the form of a *survey*--literally, "overview") means that information is gathered on a sample (often fairly large) at one point in time. For example, we might be interested in the following questions.

1. What is the rate of mental hospitalization (e.g., the number of diagnosed cases or treated cases/10,000 persons in a population) for each county in North Carolina? Let us *categorize* counties by average income level, or by degree of economic stability, or by the number of churches/1000 members of the population, or by the percentage of the population that regularly attend church ("religiosity"?)--and then determine if the rates of mental illness differ from county to county. And then, let us focus at the county level, and divide each county into areas that differ by average income, or economic stability, or "religiosity," and see if the rates *vary* as a function of our *predictor* (independent) *variables* both between counties and within counties. (If so, that would be pretty powerful evidence in favor of the effects of these variables.)

2. What is the rate of breast cancer in women (i.e., the prevalence)? In addition (looking for possible causal relationships), is there some *correlation* between predictor (independent) variables such as age, diet (e.g., fat--as a proportion of total calaries, or total fat grams/day), and/or the number of female relatives who have had cancer) and the rate of breast cancer (dependent variable)?

To answer the questions in "2" above, we might study a quota sample of 8,000 women using a questionnaire--1000 women in each age cohort from ages 10-20 through 70-80 years.

First, we determine the *overall rate (or prevalence)* of breast cancer in the sample; i.e., the proportion of the sample of 8000 that now have or did have breast cancer. Our second analysis might be more specific; e.g., the rates of cancer as a function of one independent variable, say age (i.e., rates calculated separately for each age group). This simply involves creating piles of questionnaires (subsamples)--one pile for each age group. We then examine the questionnaires within and across each age group. In reference to the overall rate, is the rate higher or lower in some age groups? If so, *do any other independent variables co-vary with age and with the rate of cancer*? That is, if cancer rates rise in the 50-60 year old category, do other independent variables concomitantly increase or decrease in that age group, in contrast to age groups with lower rates? For example, if there is enough information on the questionnaires (or if it is

generally known to be the case), we might ask if the cancer rate increases in the 50-60 year old category at about the same time as or shortly after estrogen levels decrease in that age category, or if death of a spouse is higher in that category, or if life-time exposure to radiation is especially high in that category. If two or three other independent variables *co-vary* with each other and with the dependent variable, we have a rather complex configuration of relationships. In essence, *what causes what?* To try to answer this question, we continue with the analysis, as follows.

Third, we might do an analysis (i.e., rearrange the piles--*subsamples*--of questionnaires) according to other independent variables, such as fat in the diet. For example, we could take the original 8000 questionnaires and re-arrange them in three piles (low, moderate, and high fat diets). Each pile is now composed of women with a range of ages, but similar in fat. The question, now, is *whether the cancer rate varies as a function of fat in diet*. For example, does the high-fat pile have a higher cancer rate? Now the variable--age--varies widely and similarly within each pile. Therefore, if the rate of cancer is significantly higher in the high-fat pile, and significant lower in the low-fat pile, then fat (and not age) is a major predictor (and perhaps cause) of cancer. *In other words, it seemed that increasing age was a cause of cancer, but in fact the amount of fat in the diet might increase (and exercise or other metabolic processes to reduce fat might decrease) along with age*. If so, then the apparent causal relationship between increasing age and increasing cancer rate was a spurious relationship (a spurious correlation). Draw arrows that depict possible interrelationships among the variables.

Increasing age Increasing rate of cancer

Decreasing exercise or
decreasing metabolic processes
hat handle fat

Increasing fat in diet or percentage of body fat

Fourth, we might do an analysis (i.e., re-arrange the 8000 questionnaires) on the basis of the number of women in the family who have or who had cancer. For example, there might be a pile for those "subjects" who have had no cancer in their family; another pile for subjects with one other family member who had cancer, etc. Is the rate of cancer (the percentage of persons in a pile) increasingly higher in those piles in which the number of other members in the family who had cancer is larger?

Note that you could analyze cancer histories even more precisely. For instance, you could see if the *closeness* of the biological relationship (identical twin, non-twin sibling, mother, aunt,

grandmother, cousin) predicts cancer. For example, is the rate of subjects in the pile of questionnaires that include women with an identical twin who had cancer higher than the rate of cancer in subjects in the pile of questionnaires that contain women whose mother had cancer, etc?

Finally, we might be interested in *how variables interact.* For example, we could begin by making piles on the basis of age groups. Then, *within* each age-group pile (subsample) we make smaller piles based on the percentage of fat in the diet or the percentage of body fat (say, high, moderate, and low). Then, within each age-fat pile we make still smaller piles based on the number of family members who have had cancer. The question is *which of the age-fat piles has the highest and lowest rates?* And if family history is added (i.e., we compare piles with the *same* age-fat characteristics but which differ in how many relatives had cancer), does the rate of cancer jump a great deal when more relatives had cancer, even if age-fat are the same? [Mill's Method of Difference]

A major problem with the above *cross-sectional survey research* is that many women who might have been in the sample were not--because they died of breast cancer. It may well be that something about them was different from the women in the sample who recovered from cancer. Perhaps income and education are associated with diet, smoking and environmental toxins (causes?) or with early detection and quality of treatment. In other words, the sample of 8000 may not represent the population.

B. Longitudinal

A cross-sectional study can help to reveal relationships (and nonrelationships) among variables that describe a sample. For example, we might find that the rate of breast cancer is the same across different ethnic groups; i.e., the odds of correctly predicting that a woman had breast cancer do not increase above the average chances of having breast cancer (the overall prevalence rate) if you know her ethnicity. However, *a cross-sectional study can say very little about processes or changes through time.* This is because you are *not observing persons through time*.

To better understand processes and change, longitudinal studies are useful. There are two kinds of longitudinal studies: cohort studies and panel studies.

1. Cohort studies involve examining (e.g., surveying) groups that differ by age or by location in a temporal process. For example, we might be interested in the phenomenon of "adaptation to life in a total institution." Therefore, at one time, we might study one group of patients (cohort) who are newly admitted to a mental hospital; another group (cohort) who have been hospitalized for two weeks; another group hospitalized 4-6 months; another group hospitalized

for a year; etc. Comparing the cohorts (who appear to be in different "places" in a temporal process) may reveal some kind of transition; e.g., increasing acceptance of the status of a mental patient. Similarly, at one time, we might study cohorts (different groups) of teachers: 1) those just starting their first year on the job; 2) those just starting their second year on the job; etc. By comparing the cohorts, we might find that idealism decreases during the first five years, and then rises again.

Note that a cohort study is not necessarily done through real time. Rather, the samples (cohorts) studied at the same time are at *different points in some temporal process* (e.g., the trajectory of a teaching career). Therefore, we obtain an indirect sense of changes over time.

There are at least two weaknesses in cohort studies. First, something else--*extraneous variables* (i.e., not part of being a teacher)--may *differentially* affect some of the cohorts. For example, in our study it appeared that increasingly more teachers were less idealist between years one and four. But remember that teachers in the four-years-of-teaching cohort started teaching four years ago. Perhaps something else was happening during those four years (e.g., increasing inflation) that had not happened to the brand new teachers. It could be increasing inflation (and hence a lower standard of living) that accounts for all or part of the decline in idealism.

A second weakness is that *different cohorts could be subject to differential patterns of staying in or leaving a cohort*, and these differences in staying or leaving patterns somehow *bias the composition of the cohorts* in such a way that you find what appear to changes over time, but which are really differences in leaving vs staying. For example, it could be that the teachers who began to lose their idealism quit teaching after the fourth year. Therefore, *only the more idealistic or persistent teachers remained*, and were available for study in the five-years-of-teaching cohort. In other words, *the five-years-of-teaching cohort is artificially loaded with idealistic teachers*, not because these teachers became more idealist, but because the less idealist teachers dropped out. However, a longitudinal study of the *panel*variety helps to prevent these two problems (i.e., some kind of bias in the composition of the cohorts, and some extraneous variables accounting for all or part of the apparent changes in the cohorts).

2. Panel studies involve re-studying the same sample (panel) again and again through time. For example, we might have couples estimate the amount of mutually rewarding interaction in their relationship and the amount of trust. If we have couples do these estimations each year for 5 years, we can see if changes in one variable are associated with changes in the other variable over time.