Figure 9.10. OBSERVATIONAL PLANS: The Question of Causation

Program 11 in: *Against All Odds: Inside Statistics*

This program is centred on the very important fact that a statistical association between two variables need *not* indicate that there is a cause and effect relationship between the variables. Because the different kinds of relationships that may acount for an observed association between two variables are clearest for *categorical* (*e.g.*, nominal or ordinal) variables, the program begins by describing association between categorical variables.

The relationship between two categorical variables is described by a *two-way table* of counts or percentages. Two-way tables are often used to summarize large amounts of data by grouping outcomes into categories. A two-way table allows us to calculate the *marginal distribution* of each variable alone from the row sums or column sums. We can also obtain the *conditional distribution* of one variable given a specific level of the other, by considering table entries as proportions of their row or column sums. A *segmented bar graph* visually compares a set of conditional distributions. Relationships among *three* categorical variables are described by a *three-way table*, which is *displayed* as separate two-way tables for each level of the third variable.

Three-way tables allow us to see how a third variable (a *lurking variable*) can influence the association between two variables. A comparison between two variables that holds for *each* level of a third variable can be *changed or even reversed* when the data are aggregated by summing over all levels of the third variable. *Simpson's paradox* refers to the reversal of a comparison by aggregation. In the video, Simpson's paradox is illustrated by admissions data at a fictional university: Business and Law each admit a higher percentage of female applicants than of male applicants, but the two professional schools *together* admit a higher percentage of *males*. Thus, an apparent preference for men in the overall data turns into a preference for women in each school individually.

An observed association between two variables can be due to any of *causation*, *common response* or *confounding*. Both common response and confounding involve the effect of other variables on the response. That an association is due to causation is best established by an *experiment* in which the explanatory variable is directly changed and other influences on the response are controlled. In the absence of evidence from such an experiment, causation should only be cautiously accepted. Good evidence for causation requires that the association be observed in many varied studies, that examination of the effects of other variables not remove the association, and that a clear explanation for the alleged causation exist. As an example of the way in which evidence for causation is gathered when experiments can*not* be done, the video presents a historical documentary on tobacco smoking and lung cancer. It is now accepted that smoking is a cause of lung cancer, but reaching this conclusion took many years of research.

1995-04-20

Blank page