what year is it: 2009, 5769, or some other date? The answer is that the date depends on the calendar used. There is no absolute date because there is no true zero point.

An example that illustrates both the interval level and the ratio level is temperature. On both the Fahrenheit and Celsius scales, temperature is an interval-level variable. Temperatures can be negative, and the two scales have different zero points, so that 20 °C = 58 °F, −20 °C = −4 °F, etc. After these scales had come into wide use, physicists found that there was an absolute zero temperature, than which nothing can be colder. Temperature relative to this absolute zero is measured in kelvins (K), units the same size as Celsius degrees. Measured this way, temperature is a ratio-level variable. For example, an object contains twice as much heat energy at 600 K (= 326.85 °C) as it does at 300 K (= 26.85 °C).

Numerical (ratio-level) variables in our Colleges data set include *tuition* and *enrollment*.[2]

Creating an Index with Numerical Variables

Many research applications in sociology call for the use of variables that are derived from raw sample data, but that refer to characteristics that combine information about two or more of the original variables. These derived variables are known as "indices" (singular, "index") and "scales." Although the two terms are sometimes used interchangeably, there are important technical differences between them (for a good, clear discussion and illustration of these differences, see Babbie, 2005: Ch. 6). Here we consider only the former: indices.

One of the best-known indices in sociology is the Socioeconomic Status, or SES, Index, used to measure the position occupied by an individual or group in a social hierarchy. We are well aware that income is a major factor in determining one's position, but it is also true that other factors, such as level of education, play a role. To account for this fact, we create an index consisting of information about income *and* education (and possibly other variables) and assign a score on this index to each individual or group in a sample. An easy way, although not the best way, would be to add the annual income, say in thousands of dollars, to the years of education completed. We might then divide by 2 to show that the index value is the (unweighted) average of income and education. Table 3.1 shows the procedure for a sample of three individuals.

In creating the index shown in table 3.1, we have produced additional and useful information about each individual in the sample without collecting any new data. The index is not only a convenient and more inclusive way of indicating one's status; it also gives us a different perspective on the relative positions in the hierarchy. Note, for example, that the difference between the incomes of individuals A and B is $17,000, and the difference in their educational attainments is 1 year; their SES index scores differ by 8 points. That is, A ranks above B in income, below B in education, but above B in overall SES.

TABLE 3.1 SES index, first approach

Person	*Income in thousands* *$(I)*	*Education in years (E)*	*SES index* SES = $(I + E)/2$
A	65	15	40
B	48	16	32
C	112	20	66

In Chapter 8, Box 8.1, we discuss and apply the concept of educational diversity for the states of the United States. For each state, we create an index that we call the "diversity level." This uses data on the percentage of minority students enrolled in higher education (P1) and the percentage of minorities in the general population (P2). The index is calculated by simply subtracting the percentage in the population from the percentage of students:

$$\text{Diversity level} = (P1 - P2).$$

An index value of 0 indicates that a state has met the standard of equal percentages of minorities in the schools and in the population. A positive score indicates that the standard has been surpassed, and a negative score indicate that the state has fallen below the standard.

For example, one of the states in the sample, Maryland, has P1 = 37.6 and P2 = 36.0. So its diversity level is 37.6 – 36.0 = +1.6; it has surpassed the standard. In Chapter 8, this index is used to illustrate the use of samples in the study of diversity.

Comparing Levels of Measurement

You have probably noticed at this point that some variables can be measured at more than one level. Formally speaking, this possibility arises because of the differences in the *scale* or the units of measurement we employ.[3] In the income illustration we measured the variable in words ("mine," "yours"), rank order (1st, 2nd), interval numbers (dollars above a given starting point), and ratio numbers (dollars above zero).

Researchers are interested in such a property because, ideally, we would like to use variables at the highest level of measurement possible. As we increase the level from nominal to ordinal and so forth, a wider range of statistical techniques apply. If all variables were at the ratio level, then the entire set of statistical tools could be employed in every research situation. But this is not possible in any science, and in sociology it is especially unrealistic.

In general, a variable at a given level of measurement can be "reduced" to lower levels. A ratio-level variable can be converted to an interval-level variable by shifting the zero point from an absolute value (e.g., total income) to a relative one (e. g., income

above the poverty line). An interval-level variable can be converted to an ordinal-level variable by indicating the rank ordering (1st and 2nd, higher and lower). And an ordinal-level variable can be turned into a nominal-level variable by simply designating labels with no order implied ("mine" and "yours").

This principle usually does not apply in the other direction. For example, the variable "gender" has two attributes: male and female. There is no suggestion that one is "above" the other, or that one is more than or a given proportion of the other. It is simply a nominal-level variable and must be treated as such. We can say that male incomes are higher than female incomes, but the variable of interest in such a case is income and not gender. The same is true of most other nominal-level variables: student status (registered or not registered), residency, religious affiliation, and countless others.

Social researchers do make a special, widespread exception to this rule. This is in the case of treating a certain kind of ordinal-level variable as if it were interval-level. The type of variable is familiar to anyone who has participated in a social survey, and it is known as a Likert scale (the first syllable rhymes with "pick"), named after its inventor, Rensis Likert. A Likert scale item presents to the respondent a stimulus, in the form of a statement or phrase. The respondent is then instructed to select one of a set of answers that reflect his or her knowledge, attitudes, or practices and that have a clear ranking. An example of a Likert item and some hypothetical responses are shown in Box 3.2.

Because a Likert scale is clearly an ordinal-level variable, we can easily compare respondents by ranking them on each item. Let us say that you checked "I strongly

BOX 3.2 Statistics for Sociologists

Likert Scale: Ordinal or Interval?

The Board of the University has proposed dropping the football program. Indicate how you feel about this (check one).

☐ I strongly disagree	☐ I disagree	☐ I have no opinion	☐ I agree	☐ I strongly agree
1	2	3	4	5

Outcome

			interval a	*interval b*
You	"strongly disagree"	I	You	Other
I	"have no opinion"	1	3	5
Other	"strongly agrees"			

Does *interval a* equal *interval b*?

disagree," and I checked "I have no opinion." Then it is obvious that I rank below you (moving from left to right) in our feelings toward the proposal: you are more opposed than I am. Now, suppose that we learn of the response of another person, which is "I strongly agree." That person indicates even less disagreement than I do, and among us the other can be ranked lowest, with me in the middle and you the highest. Moreover, we can say that the other person is three positions below me and four below you. But can we say that there is a two-point difference between you and me, a two-point difference between the other person and me, and a four-point difference between you and the other person? If so, then we could also say that the difference (or interval) between you and me is the same as the difference between the third person and me. That is, we would be able to treat an ordinal-level variable as if it were interval-level. So, for instance, we could use the numbers 1, 3, and 5 to stand for the three responses, add them up (1 + 3 + 5 = 9), divide by 3 (9/3 = 3), and refer to 3 as the average of our responses.

In order make this kind of judgment, some basis must be available to indicate that the difference in the intensity of feeling between any two points on the scale, say between "strongly agree" and "agree," is the same as that between any other, say between "disagree" and "strongly disagree." But this is usually not known; nor is an attempt ordinarily made to establish such a basis. Nevertheless, researchers routinely interpret the numbers associated with Likert items as if they referred to amounts. This does create misleading results in some instances, but the problem is not considered to be serious (see Kim, 1971; 1975; Smith, 1978).

Independent and Dependent Variables

Up to this point we have focused on hypotheses with only one variable in their predicate parts, such as a crime rate, income, or response to a Likert scale item. Statistics and statistical techniques that refer to these are, for obvious reasons, called *univariate.* Although much social research is of the univariate kind, it is far more common to encounter studies of units for which data on more than one variable are employed. Hypotheses and statistical techniques that include two variables are, for equally obvious reasons, known as *bivariate,* whereas applications with three or more variables are called *multivariate.* For the most part, the descriptive and inductive statistics featured in this book are of the first two types: univariate and bivariate. Only toward the end of Chapters 4 and 11 through 14 do we briefly consider multivariate approaches.

Among the several concerns associated with using more than one variable, none is more important than the *relationship(s) between and/or among the variables*—an issue that, of course, does not arise with only one variable. In bivariate applications, the researcher must decide which of the two variables is *independent* and which is *dependent.* This decision is usually based on the theory from which the research

hypothesis is drawn; that is, if past research has shown or assumed that a particular variable is dependent on another, then current researchers follow that guideline (until observations indicate otherwise). As a general rule, especially when the relevant theory is not known, we say that any change(s) or difference(s) in the attributes of the dependent variable are the result of changes in the independent. This can be symbolized as:

$$\text{Independent} \rightarrow \text{Dependent}$$

Thus, when deciding between variables A and B, we ask whether it makes more sense to say that changes in A depend on B or that changes in B depend on A. That is,

$$B \rightarrow A \text{ or } A \rightarrow B?$$

In the first case B would be the independent; in the second case the independent would be A. An equivalent way to phrase this is that the independent is a "probable cause" of the dependent—although it is not necessary for our hypotheses to imply cause and effect, just "association."

Whether or not a characteristic to which a particular variable refers actually is the cause of another (e.g., whether variation in level of education is the cause of variation in income) is a difficult matter to determine. At least three things must be established before we can conclude that a relationship is causal:

1. The independent variable must precede the dependent in time.
2. The two variables must be clearly associated, in that they vary together in an observed manner (e.g., high education level goes with high income and low education level goes with low income).
3. All other possible causes, direct and indirect, have been eliminated.

None of these criteria is easy to establish, but one can *never* be certain about the third. Thus, the search for newly discovered causes of observed phenomena is a major undertaking that makes science a never-ending enterprise.

The procedure for determining *probable* cause is far less demanding. It can be illustrated with two of the variables already discussed: income and response to the "football" Likert scale item. Suppose (1) your income is higher than mine and (2) you agree with the proposition to end football less than I do. Which (if either) is the more likely possibility: that our difference in income is the result of our difference in attitude toward the proposal, or that our difference in attitude is the result of our income difference? Although the decision is not always so clear, it seems obvious in this illustration that it is the second option—our incomes are affecting our attitudes.

Two points need to be stressed here. First, no variable is, in itself, either independent or dependent. This is always decided by virtue of its theoretical or logical

relationship to another variable. For example, suppose that we were considering the variables (1) level of education, as measured by years of school completed, and (2) income in dollars. Now at first thought it would seem obvious that education is the independent variable in this pair. After all, don't we improve our level of education to increase our earning power, and thus our income? This is quite true. But if we were to specify further that by "income" we meant income of our parents and by "education" we meant our own educational level, it would be clear that income is the independent variable: wealthy parents can pay more to educate their children. It is the context that makes a variable independent or dependent, not the variable itself.

The second point to be stressed is that in nearly all bivariate applications, the independent and dependent must be specified, even if there is no clear choice. In such instances, the researcher may need to guess—knowing that it may well be a wrong guess—or make a choice based on convenience. Otherwise, it is usually not possible to proceed with the analysis. Bivariate tables require one to indicate which variable's attributes are shown in the columns (the independent) and which in the rows (dependent). Bivariate graphs require one to indicate which variable is represented along the horizontal or x axis (independent) and which along the vertical or y axis (dependent).

In multivariate applications, the choices become a little more complex. Ordinarily, we have a fairly good idea (from theory and context) of what the dependent variable should be, although sometimes even this is unclear. But even if the dependent variable can be established, we still must determine how to treat the other two, three, or more variables; and there are several possibilities, some of which are shown in figure 3.2. If we assume that the dependent variable, C, is response to the football proposal, that variable A is education, and that variable B is income, then each part of the figure would be interpreted as follows:

1. Both education and income are independent variables because they affect one's response, but they do not affect one another.
2. Education affects income, and income affects one's response. In this case, income is an *intervening* variable (because it intervenes between the independent and dependent variables).
3. Education affects income, education affects one's response, and income separately affects one's response. In this case, income acts as both an independent and an intervening variable.

We briefly discuss these and other multivariate possibilities, as well as techniques for describing and generalizing about the relationships, in Chapter 4.

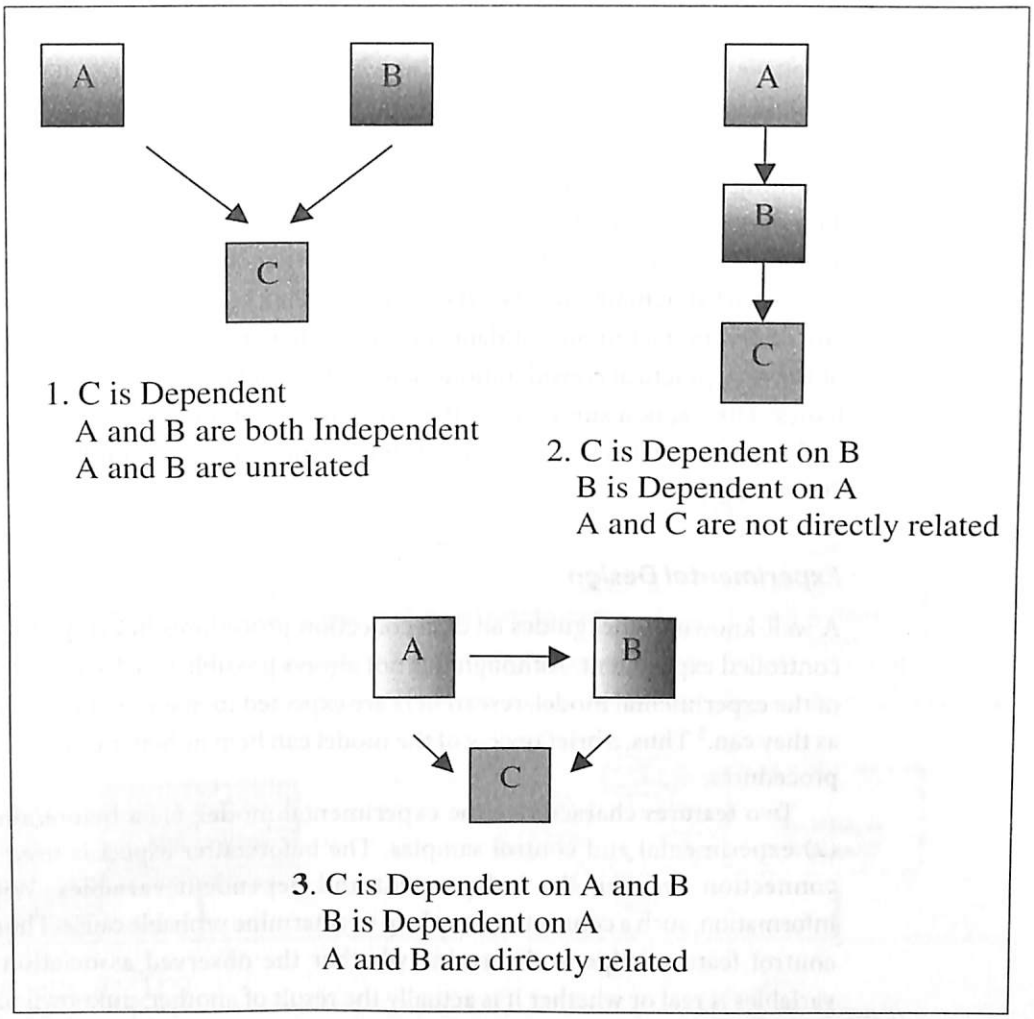

FIGURE 3.2 Three types of multivariate relationships.

Where Do Social Scientific Data Come From?

This section discusses the various sources of data used in social research. We begin with a brief summary of the most common techniques of data collection, including the approach that many people equate with social science: survey research. Next, we introduce several useful sources of data that have already been collected and are available online and/or in printed versions. In the following section, we use some of these data to illustrate basic procedures in descriptive and inductive statistics.

Collecting Data

When we refer to samples and units as what we *observe*, we mean more than just casually taking notice, as in "I observed the color of your car." Instead, to observe in the statistical sense involves carefully noting one or more characteristics of each unit and recording what our senses tell us. This act transforms our sense impressions into "raw" data that can be used for descriptive or inductive purposes. Thus, the term *data collection* is synonymous with *scientific observation*, and it more precisely conveys the sense of what actually occurs.[4] As is true in other scientific fields, sociologists employ any of several techniques of data collection, depending on the nature of the problem of interest, practical considerations, and—of special concern in social science—ethical issues. This section summarizes the five major techniques of data collection in sociology: experimental design, surveys, ethnographic research, secondary data analysis, and content analysis.

Experimental Design

A well-known model guides all data collection procedures in every scientific field: the controlled experiment. Although it is not always possible to satisfy all of the conditions of the experimental model, researchers are expected to make as close an approximation as they can.[5] Thus, a brief review of the model can help us better understand most other procedures.

Two features characterize the experimental model: (1) a before/after design and (2) experimental and control samples. The before/after aspect is used to establish a connection between the independent and dependent variables. With additional information, such a connection may help to determine probable cause. The experimental/control feature helps to determine whether the observed association between the variables is real or whether it is actually the result of another, unknown or unmeasured variable. Such a situation, in which the observed association is not authentic, is referred to as a *spurious* (literally, "counterfeit") relationship.

In the illustration shown in figure 3.3, we are testing a hypothesis that stipulates: If work teams watch an instructional film about avoiding errors in the production of microchips, then the performance of the teams will improve. Here the unit of observation (and analysis) is the work team, the independent variable is watching the film, with the attributes "yes" and "no," and the dependent variable is error reduction. The dependent can be measured in any of four ways.

- Nominal: "yes improved" and "no did not improve"
- Ordinal: the teams that do watch the film improved "more" or "less" than the teams that do not watch the film

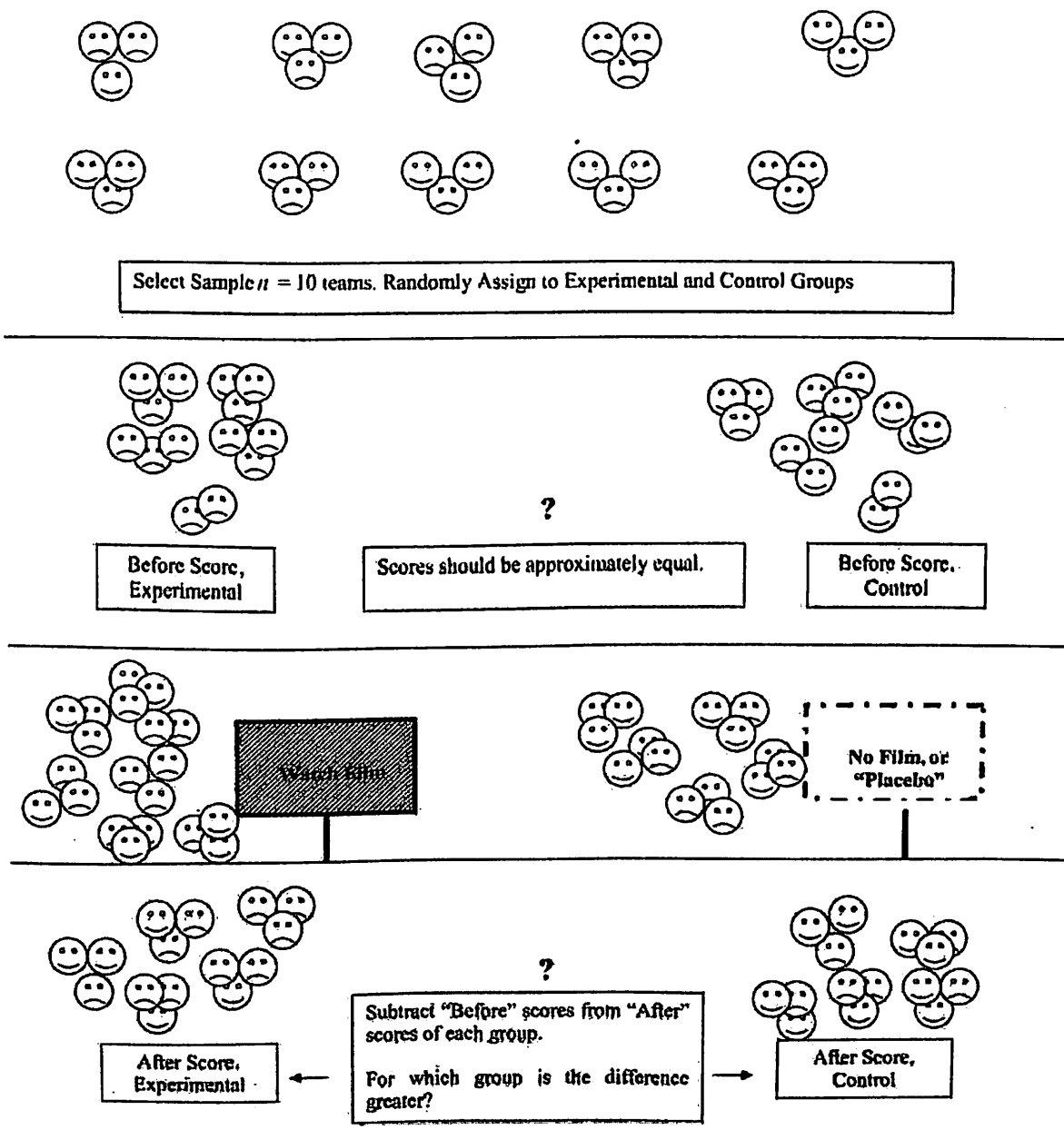

FIGURE 3.3 The experimental model.

- Interval: measured in the difference in the number of errors above a given level—such as above last week's average
- Ratio, difference in the number of errors after and before the film, from zero on up

For the experiment, 10 work teams of three persons each have been selected from a population of all work teams in a manufacturing facility. With this completed, the sample is divided into two equal-sized subsamples, each with $n = 5$, one of which is to be the experimental group and the other the control group. It is important to ensure that the teams are unaware of the group to which they have been assigned. Also, the two subsamples are to be as much alike as possible so that no difference between them exists that would influence the experimental outcome. This is ordinarily achieved through random assignment, such that chance and only chance determines whether a specific team is placed in one or the other of the two groups.

Next, data are gathered about the performance of each of the teams. This is done so that the experimental and the control group can be given a "before" score, either the total number of errors or an average for each team. At this point, the independent variable (watching the film) can be introduced. Only the experimental group actually watches the instructional material. The control group either watches an entirely unrelated film (such as a documentary on the Grand Canyon) or it watches none at all. If used, the unrelated film plays the role of a *placebo*, the "sugar pill" of medical experiments. This leads those in the control group to believe that they are being "treated," and it preserves the lack of knowledge about which of the groups is the control.

Finally, after the instructional film is shown to the experimental group, each group's performance is tested again. This is done to establish a pair of "after" scores. If the number of errors decreases in the experimental group and either fails to decrease or decreases considerably less in the control group, the researchers can conclude that the hypothesis is supported—or at least that they are on the right track. Any other outcome—for instance, that neither group experiences a decrease—signals that something is wrong, including that the hypothesis does not stand up.

It is important to note that groups, rather than individuals, are being compared. The reasons are discussed further in Box 3.3.

The principal advantage of the experimental model is contained in the name of one of the groups: *control*. The researchers can select their sample, assign units to one or the other group, control how and when the independent variable will enter into the research, measure the dependent variable twice, and rearrange the procedure at will until it is adequate for their purposes. These factors make it the ideal model for scientific data collection.

However, there are several disadvantages. Perhaps the most obvious is that severe legal and ethical constraints apply to experimentation with human subjects. All such

BOX 3.3 Statistics for Sociologists

The Focus on Aggregates

In our discussion of the origins of sociology and statistics in Chapter 1, we stressed that the principal focus of sociological research is not on individuals but on aggregates. You may be familiar with one of the true classic discussions of sociology's concern with aggregates, *The Rules of Sociological Method* by Émile Durkheim (1982). In this book, Durkheim makes a strong case for "social realism," the view that aggregates have a true concrete existence beyond the individuals who make them up. At the same time, he argues against the practice of "reductionism," the attempt to explain aggregate characteristics solely on the basis of individual characteristics—an application of the saying that the whole is always greater than the sum of its parts. This is also the context in which Durkheim introduces the now-famous concept of *social facts*.

Aggregates can range from a two-person relationship, such as a pair of friends, all the way to the entire population of the world, approximately 6.5 billion people. Of course, it is often the case that the units of observation in a research project are individuals. But, as Durkheim pointed out, the sociologist is not concerned with their unique characteristics but rather with the larger categories to which they belong: gender, social class, membership in a particular organization, and so on.

The example of the experimental approach illustrates a common way in which aggregates are treated as units of observation. You will note that the units are work teams (each consisting of three individuals), and it is their performance as teams that is of interest. The data that are collected apply to the teams, not to the individual members; and the variables, watching a film (the independent) and performance (the dependent), are characteristics of the teams as well.

In experimental designs such as this, individual as opposed to aggregate characteristics are not of interest. In fact, procedures are ordinarily employed to nullify or hold constant individual differences. The most common, and typically the most effective, of these is known as "randomization" or "random assignment." With this procedure, individuals involved in the experiment are assigned to a specific aggregate (one of the work groups in our example) based on chance alone. This might involve the toss of a coin or selection of a random number. In this way, an attempt is made to minimize the probability that the assignment is based on whether one is male or female, young or old, skilled or unskilled, etc. In other words, the aggregate's performance is emphasized and the performance of the individuals is discounted.