# CHAPTER 2
# SOME BASIC CONCEPTS

## Outline


Key Terms
- *Variables*
- *Levels of measurement*
- *Tools for working with measurement levels*
- *Scaling*
- *Other terms*
- *Labeling variables*
- *Constants*
- *Functions*

Units of Analysis

Data Structure

Sample Size and Sample Selection

What Have We Learned?

Applications

Exercises

In this chapter we will define a few basic concepts that are necessary before we go on to examine data using models. You may already have a common sense understanding of these concepts or have encountered them in other courses. But we want to clarify what these concepts mean as they are used in quantitative analysis. We will reinforce the discussions of chapter 1 by paying special attention to the concepts that are needed to understand models and how we apply data to evaluating models.

## Key Terms

We have to define a number of terms before we can proceed. While at first this seems a dull enterprise, we can enliven it a bit by keeping in mind that in defining terms we are constructing a way to look at the world. Our terms generate a conceptual framework, and in the case of statistics and quantitative methods it is a framework that has substantial power and subtlety.

### Variables

Variables are in some fundamental sense what we are trying to understand. They are the properties or characteristics of people or organizations or nations that we want to explain. In particular, we want to know why people, organizations, or nations are different from one another – we want to explain variation. We want to know why the characteristics (the variables) vary. So variables are characteristics, such as gender or years of formal education for an individual or level of affluence for a nation.

A variable must vary – across individuals or countries or organizations it must take on different values. One can distinguish disciplines by what varies in what they study and what does not. For example, the species under study does not vary for most social scientists – we usually study only human beings. So species is not a variable for most social scientists. But zoologists and some biological anthropologists and psychologists are concerned primarily with differences across species, so for these researchers, species is a variable.

Variables have attributes that describe the variation in the characteristics they measure. For the variable "gender" the attributes are male and female. For the variable "years of formal education" the attributes are 0, 1, 2, 3, and so on. For the poverty rate for a state, the rate can range from 0 percent, when no one is poor, to 100 percent, if everyone is poor. Of course, we may not have any states at 0 percent or at 100 percent or at any other particular value. But the variable *might* take on those values.

### Levels of measurement

The different ways we collect data lead to measurements with different characteristics. In quantitative work we always assign numbers to what we observe in the

**Box 2.1 Keeping Percentages and Proportions Straight**

The poverty rate is calculated by counting the number of people in poverty. Or, if we can't count the number of poor people, we can use surveys and the statistical tools that apply to them to estimate the number of people in poverty. Then the number in poverty is divided by the total population. This yields a number between 0 and 1.00. This is referred to as the *proportion*. For convenience we often change this into a *percentage*. A percentage runs between 0 percent and 100 percent. Remember that "percent" is Latin for "divided by 100." So 50 percent is not 50, it is 50/100 or 0.50. For most people it's easer to read 50 percent than 0.50, but remember that 50 percent is not the same as 50. It's the same as 0.50. The proportion is calculated by dividing the size of the group of special interest, the number in poverty in this case, by the total population. The percentage is just the proportion multiplied by 100. While this distinction may seem minor, it becomes important when we do arithmetic with either proportions or percentages, as we will in later chapters. If we use percentages when we should use proportions we will get the wrong answers.

Sometimes we use percentages and proportions to examine how much something has changed. For example, we might have data on how much the homicide rate changed over 10 years. When the proportion or percentage is calculated based on how many people (or countries or states) have a certain characteristic compared to the population, then the proportion must be between 0 and 1.0 and the percentage between 0 and 100. But if we look at change these restrictions don't apply. Suppose the number of homicides went from 50 to 150 over a few years. The change would be the new number of homicides (150) minus the old number (50), which is an increase in homicides of 100. We then divide by the original number (50) to get the *proportional change* (100/50 = 2.0). This means that the proportional change is 2. If we want the percentage change we would multiply by 100 and get a 200% increase in homicides. So when we look at change, there are no limits to the numerical values we can get for proportions and percentages.

By the way, when looking at proportional or percentage changes, be sure to keep in mind the base. A change from 50 to 150 seems like an important change. But a change from 5 to 15 is also a 200 percent increase. You have to decide what changes are important. Percentages and proportions help but you must use judgment too.

world, following the idea expressed by the physicist William Thomson in the quote in Box 2.2. We can have different kinds of information expressed by the numbers we assign: in some cases they are just labels, like numbers on an athletic jersey. In other cases they are the "ordinary" numbers of arithmetic. The amount of information in a number we use as a measurement is called the **level of measurement.** For

**Box 2.2**

> I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind: it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of *science*, whatever the matter might be. (William Thomson (Lord Kelvin), *Popular Lectures and Addresses* (p. 80), from Gaither and Cavazos-Gaither, p. 139)

This is a rather extreme position. We would agree that it is often useful to express things in numbers, but there are useful ways of doing science without numbers too.

some kinds of numbers, the standard tools of arithmetic apply, for others they do not. This in turn determines what statistical techniques can be applied to variables.

Most researchers refer to four levels of measurement: **nominal, ordinal, interval** and **ratio**. The idea of four levels of measurement comes from a famous paper by the psychologist S. S. Stevens (Stevens, 1946). His typology is used in most statistics books, and many researchers feel we should be very cautious about using only statistical methods that match the level of measurement. But many other researchers, including us, are less strict about this. A good review of the controversy can be found in Velleman and Wilkinson (Velleman and Wilkinson, 1993). We will explain the controversy after we describe the levels of measurement.

*Nominal*

In *nominal* measurement we simply assign observations to categories, and there is no ordering among the categories. One example would be student ID numbers – each person in a study gets a different number so their information can be identified. Another example is gender, which we can usually think of as having two attributes: male and female. These attributes have no special numerical value and cannot be rank ordered in any meaningful way. Thus, gender is a nominal level variable.

Sometimes numbers are given to the nominal categories of a variable, for example 1 = female and 2 = male, because numbers are more convenient in computer programs used to do statistical analysis. The numbers do not imply rank ordering, but are arbitrary. We could just as well make male equal to 1 and female equal to 2. The numbers assigned don't matter (as long as we keep track of them, so we don't confuse categories). Thus we might assign numbers for marital status as follows:

1 Married;
2 Widowed;
3 Divorced;
4 Separated but married; and
5 Never married.

Or for continents we might assign them as:

1 Africa;
2 Antartica;
3 Asia;
4 Australia;
5 Europe;
6 North America; and
7 South America.

A nominal variable should assign every person (or country or state) to a category, and to only one category. We think the marital status variable does this – everyone could be assigned into a category and no one would fit into more than one category. The continent variable works only if we leave out the many island nations that are not traditionally "assigned" to a category. And we also have to make a

## Box 2.3 Historical Point about Assigning Categories to Variables

Developing procedures for assigning people to categories in a categorical variable sometimes has important policy implications. In the US, every Census since the first in 1790 has included a racial classification variable, but the categories used in this nominal variable have changed over time, and the racial categories were always related to public policy (Hattam, 2005; Prewitt, 2005; Snipp, 2003). In the first Census in 1790 there were three categories: "European," "African," and "Indians not taxed." As the phrase indicates, "Indians not taxed" were not counted for the purposes of representation in Congress, while for the purposes of allocating seats in Congress, a slave counted as "three fifths" of a freeman. The federal marshals who conducted the Census made the determination of who fitted into each category. The assumption of the time was that for people with parents, grandparents, and other ancestors of African heritage, the "African" group "dominates" in assigning an individual to one category in the nominal variable. This is called the "hypodescendent presumption" or the "one drop rule" in which any African ancestry led an individual to be classified as of "African" descent (Hollinger, 2005). Such a rule obviously does not reflect the reality of race and ethnicity in the US or elsewhere. Starting in 2000, the US Census used five primary categories: "American Indian/Native Alaskan," "Asian," "Black/African-American," "Native Hawaiian/Pacific Islander," and "White" plus "Some Other." To get past the hypodescendent presumption, respondents to were allowed to "select one or more." This leads to a lot of categories once one looks at all the possibilities (63 categories in fact) but gets closer to the reality of race/ethnicity than the earlier, cruder distinctions.

decision about what to do with, for example, Russia, since Russia spans both Europe and Asia. We should also note that Iceland, Ireland and Great Britain are usually assigned to Europe even though they are islands. We could add a category to account for Oceania and one for the Caribbean but there are still other island nations.

With nominal levels of measurement it is not legitimate to perform arithmetic on the variables since the numbers are simply shorthand versions of words, and we cannot perform arithmetic on words. But there is a special trick that is worth mentioning. Suppose we think of a nominal variable as indicating whether an observation has some property. This works best with just two categories, like gender. We can call these binary nominal variables or just **binary variables**. Then we can create a variable that is coded 1 for women and 0 for men. These are sometimes called "dummy" variables because when they were first developed there was a sense that they were not "real" variables. The term has stuck but as we will see, binary variables are very useful and just as "real" as other variables.

The new binary variable indicates whether an individual has the property "female." It captures the same information as the 1, 2 gender variable, but it has the advantage that certain kinds of arithmetic are legitimate with the binary variable that would not make sense with nominal variables that don't have values of only 0 and 1. For example, if we take the average of the new female variable, we would add up all the 1s and 0s and then divide by the number of people in the data set. In doing the addition, each woman counts 1, and each man counts 0. So when we take the average of a binary nominal variable we are really adding up the number of women and dividing by the number of people. But if we took the average of "continent" or "marital status" variables with numbers assigned as above, the results would not make sense.

Taking the average of a binary variable is how we calculate the proportion of people in the category labeled 1 – the number of people in the category of interest divided by the total number of people. So taking the average of a 0–1 variable is the same as calculating the proportion of people in the category coded 1. For example, in the 2000 International Social Survey Programme (ISSP) data set, there were 13,964 men and 17,064 women. If we add those together we get 31,028 respondents. Then if we want the proportion of women, we would divide 17,064 by 31,028 to get 0.550 or 55.0 percent.

But if we score the men 0 and the women 1 and take the average we will add together 17,064 ones and 13,964 zeros for a total of 17,064. Then, in taking the average, we take the sum and divide by the number of observations, so we divide 17,064 by 31,028 to get 0.550, or 55.0 percent. So whether we take the proportion or the average of a "zero, one" variable, we get the same thing. But remember this *only* works for a "zero, one" variable and only gives us the proportion in the category scored one. When we have more than two categories, things get a bit more complicated, and we will save the explanation of how to deal with that situation until later.

*Ordinal*

In ordinal variables numbers represent rank ordering, like the finishing order in a race. But we do not know from them how far apart the ranks are. Many attitude scale items take this form. For example in a survey we might ask:

> Tell us how you feel about the following statement. Do you strongly agree, agree, neither agree nor disagree, disagree or strongly disagree: It is right to use animals for medical testing if it might save human lives.

We then code the responses as:[1]

1 Strongly agree;
2 Agree;
3 Neither agree nor disagree;
4 Disagree; and
5 Strongly disagree.

Here the numbers mean more than they did in a nominal scale. Given how the numbers have been assigned, we know that people coded 1 (strongly agree) feel more strongly that animal testing is alright than those coded 2 (agree) who feel more strongly than the people coded 3 (neither agree nor disagree) who feel more strongly than those coded 4 (disagree) who feel more strongly than the people coded 5 (strongly disagree). But we do not know the distance between a "strongly agree" and an "agree," nor do we know if that is the same as between any other two points on the scale. We know the rank ordering but we can't assume that the difference in feeling between a "strongly agree" and an "agree" is the same as between a "disagree" and a "strongly disagree" (even though if we consider the 1, 2, 3, 4 and 5 values as regular numbers, we would think the distances are the same). As with nominal data, we have to be careful how we use the numbers we've assigned.

Why is that? What happens if we take the average? (This is in fact something we really shouldn't do with ordinal data.) In the 2000 ISSP data set, the distribution of valid responses was as shown in Table 2.1.

We could add up a 1 for each "strongly agree," a 2 for each "agree," a 3 for each "neither agree or disagree," a 4 for each "disagree" and a 5 for each "strongly disagree." Then we would divide by 29,486, the number of respondents, to get the average of these scores.

Doing the arithmetic, we will find that the average is 2.49. But since this is ordinal data, it's not clear what that number means. We don't know that the difference in strength of feeling about animal testing, say between "strongly agree" and "agree," is exactly one point, even though that is how it's scored. Nor do we know that the difference between "agree" and "neither agree nor disagree" is also exactly one point. There has been a long debate in the statistical literature on whether we should apply techniques that require arithmetic to interval data. We will discuss this issue below.

**Table 2.1** Distribution of responses to question on animal testing

| *Response* | *Number* | *Percent* |
|---|---|---|
| 1-Strongly agree | 5,520 | 18.7 |
| 2-Agree | 12,803 | 43.4 |
| 3-Neither agree nor disagree | 4,617 | 15.7 |
| 4-Disagree | 4,161 | 14.1 |
| 5-Strongly disagree | 2,385 | 8.1 |
| Total | 29,486 | 100.0 |

*Data source*: 2000 ISSP data set, analyzed with SPSS. Details of sampling strategy and related information for continuing examples are provided in the Applications section.

*Interval*

Here we have "regular" numbers in which the number indicates how far apart observations are. Such things as homicide rate for a state or income, and years of formal education for individuals are interval observations. For all practical purposes we can perform conventional arithmetic with these numbers and get meaningful results. For example, the sum of the incomes of everyone in a city is a meaningful summation – it is the total income for everyone.

*Ratio*

Ratio numbers have all the properties of interval numbers, but also have a natural zero point. The natural zero point (having none, or zero, of the quality being measured) is important for performing division. But in practice we will not worry about this distinction and just refer to interval variables whether they have a meaningful zero point or not.

## Tools for working with measurement levels

We can always move from higher levels of measurement to lower. For example, when we have an interval number, like age, we also know the rank order of people in terms of age, so we can treat age as an ordinal number. We can label people with a particular age, so we can create a nominal variable. But we have to be careful about treating variables at lower orders of measurement as if they had the properties of higher orders of measurement. Why? Many methodologists and statisticians argue that converting ordinal variables into interval variables is not a good idea because we only know the rank orders, not the distance between scores. Standard arithmetic assumes that we know the distance between numbers. For example, the difference in age between a person who is 20 and one who is 21 is one year, and the difference in age between someone who is 42 and one who is 43 is also one year. But as we noted above, the difference on a survey response between a "strongly agree" and an "agree" may not be the same as the distance between an "agree" and a "neither agree nor

disagree." So the arithmetic may lead us astray. And with nominal variables we don't even have a rank ordering so any arithmetic is almost certain to lead us astray.

Most statistical tools have been developed for either interval or nominal variables, taking account of their different properties and the kinds of mathematics that can be applied to them. We have extensive tools for interval variables. We have a number of good tools for nominal variables and a mixture of nominal and interval variables. But we have only very limited tools for ordinal variables and mixtures of ordinal and nominal or interval variables. This is unfortunate because in the social sciences we collect a lot of data through surveys, often using attitude questions like the one above that give us interval data.

There are three ways to deal with ordinal data. One is to use the rather restricted set of tools that take account of the fact that we have only rank orderings, and we don't know the actual distances between values and thus can't do normal arithmetic. Another is to convert ordinal measures to nominal measures and use the extensive set of tools that exist for nominal measures. This is a reasonable strategy but it has some costs. First when we do this we throw away the information about the order of responses, treating them just as categories with no order. Second, for technical reasons you'll see later the techniques for analyzing nominal measures often require large sample sizes.

The third strategy is more controversial. We can pretend that ordinal measures are actually interval (this is sometimes called "scaling up"). We can proceed to do the kinds of arithmetic that aren't really justified for ordinal numbers, like adding the numbers together and dividing by the number of observations to get the average. There is some research that indicates treating ordinal measurements as if they were really interval will not lead us far astray.[2] But there are no guarantees that in any particular analysis we will get the right answers. A good strategy for this and other problems in statistical analysis is to try to do the analysis several ways and see if they all lead to the same conclusion.

## Scaling

Ordinal levels of measurement are not easy to analyze, but are very common because that's the kind of data we get from most survey questions. To get past this problem, many researchers combine several nominal or ordinal variables into a new variable called a "scale." There are many statistical tools that help us develop good scales, and scale development is a specialty for many bright researchers.

We create scales for two reasons. First, we can't do ordinary arithmetic on the individual ordinal items, but we are usually comfortable doing arithmetic with a scale, allowing us to use many techniques in analyzing the scale that we can't use on a single ordinal item.

Second, we build scales because we believe that several items in a survey get at the same value or belief or attitude of our respondents. If we combine them into a single variable, a scale, we hope we'll get a better measure of the value, belief, or attitude. This is because we think that the errors in one question will balance out

the errors in the other questions. This is a common way of dealing with measurement error. If we have several measurements and we average them, the average is likely to be a more accurate measure than any single measurement.

For example, say we had data on the following question:

> Tell us how you feel about the following statement. Do you strongly agree, agree, neither agree nor disagree, disagree or strongly disagree: Animals should have the same moral rights that human beings do.

We could combine this with the animal testing item, since we feel that both these items reflect a person's attitude regarding the position of animals and humans in the world. We hope combining the two items into a single scale would more accurately reflect people's attitudes than either item alone. Unfortunately though, this question was only asked in the American survey of the ISSP and was not available for respondents in other countries, so in this book we will have to use only the animal testing question to measure animal concern.

## Other terms

While nominal, ordinal, and interval are the classical terms used to distinguish levels of measurement, other terms are used as well. We sometimes refer to **qualitative** and **quantitative variables**. By qualitative we mean nominal variables, and quantitative refers to ordinal and interval levels of measurement.[3] Nominal variables are also called **discrete variables** (because they consist of separate, distinct categories). Interval variables are called **continuous variables** (because they can take on any possible value).[4] Ordinal variables are somewhere in between, although most authors call them discrete. If we have only a few possible scores on a variable, an ordinal variable looks discrete. This is the case for the five possibilities on the response scale for the animal rights question: "Strongly agree, agree, neither agree nor disagree, disagree, strongly disagree."

## Labeling variables

In the standard notation we use in this book, letters from the end of the alphabet will indicate variables. Recall from chapter 1 that the letters X and Y refer to variables. Thus we might label gender X, level of education Z, and attitude about animal testing Y. For states we labeled the poverty rate X and the homicide rate Y. Or sometimes we use computer abbreviations, which are just short sets of letters to stand for the variable. So gender could be "gender" or "sex" and years of education could be "education," "ed," or "educ." What approach we use doesn't matter as long as we keep track of the variables.

Remember that each variable takes on some value for every unit of analysis included in our data. If X is gender, then every person in the data set will have a score of 0

or 1 (if that is how we coded gender), and there will be as many values of gender as there are people (of course many people will share the same value).

We use numbers to label people 1, 2, 3, . . . , and so on. These are just arbitrarily assigned "ID" numbers that help the computer keep track of the data. For example, we might assign ID numbers to states after listing them in alphabetical order so that Alabama gets a 1, Alaska a 2, and Wyoming a 50. In a survey, people are assigned ID numbers either at random or as their data are entered into the computer. But however the IDs are assigned, they don't mean anything; they are just a way of keeping track of the data. For states, we could use the state names but it's somewhat harder for the computer to work with words than with numbers. In surveys, we almost always guarantee respondents confidentiality, so we don't want their names in the data set.

We also distinguish between *independent* and *dependent variables*, as noted in chapter 1. Recall that dependent variables are what we are trying to explain. We want to know why they vary from person to person, state to state or country to country. Independent variables are variables we are using to try to understand why the dependent variables vary. We might think that homicide rate is a function of poverty. Then the independent variable is the poverty level and the dependent variable is the homicide rate. We might think that attitudes towards animals are a function of gender. Then gender is the independent variable and may be one of the questions in a survey measuring *animal concern*, which is the dependent variable. Theory might suggest that differences across people, states or countries in the independent variable produce variation in the dependent variable.

## Constants

Variables vary across observations. But many statistical procedures also produce constants, which are numbers that are the same for everyone in a data set. A simple example is the average score of all members of a class on a test. Each student has her or his own score, but there is one average for the class. In that case, the average is a constant. But we might compare test scores across classes. Now our unit of analysis has shifted from individuals in a class to classes, and the average score for a class becomes a variable since it will vary across classes.

## Functions

As noted in chapter 1, a function links the independent to the dependent variable. We are going to expand our discussion of functions here because we will use the idea of a function throughout the rest of the book. Please don't worry about the algebra. Here and in the chapters that follow we'll take things slowly, one step and a time, so you won't get overwhelmed even if you find the equations daunting. Remember that in reading equations, look inside each group of parentheses first, then out to the next set, and so on.

To start, remember that we usually call X the independent variable (for example the poverty rate of a state) and Y is the dependent variable (for example the homicide rate). In this model, we would be thinking that perhaps the variability across states in the homicide rate is caused by states differing in their poverty rates. To show that idea we would link the homicide rate Y to the poverty rate X with a function. It is useful at this point to reproduce the equation from chapter 1 that illustrates the relationships between X, Y, E and f:

$$Y = f(X) + E \tag{2.1}$$

As we discussed in chapter 1, E is an error term that indicates we do not expect the poverty rate to exactly predict the homicide rate. E is a variable because the amount by which the prediction for the dependent variable misses will differ from person to person or country to country. So this simple function has three variables: Y, X, and E.

The function, f, represents how Y is related to X. For the simplest models (which are often very useful), we assume the function is a simple equation that maps values of X to values of Y in a straight line. A straight line has the form $Y = A + (B*X)$, which we read "Y equals A plus B times X." That equation defines a straight line. But when we analyze data we need to include an error term E that indicates that we don't expect the data to fall exactly on the line. So the equation that is our little model of the homicide rate for states is:

$$Y = (A + (B*X)) + E \tag{2.2}$$

Now that we have actually written down a specific function, rather than letting f(X) stand for any possible function, we have two constants A and B in addition to the three variables, Y, X, and E. There will be one A and one B for all the states. A and B describe the link between X and Y for the group of states studied. The equation says that a state's homicide rate is predicted to be:

$$(A + (B*X)) + E \tag{2.3}$$

(Remember that the parentheses tell you the order in which to do the arithmetic: you should multiple B times X first, then add A.)

In algebra our equations do not have an E term because in algebra there is a perfect relationship between X and Y. We often find such equations capture practical advice. For example, there are equations that calculate the minimum heart rate that people should achieve when exercising to get the full benefits of their workout. The minimum heart rate data is based on what is called Karbonen's formula. There is also an equation for the maximum heart rate you should strive for when exercising so as not to go too far. Of course, if you haven't been exercising, it's a good idea to check with a health care professional rather than using one of these formulas.

If we call P the minimum pulse rate you should get to when exercising and X your age, then the equation for the minimum rate for mild exercise is:

$$P = A + (B^*X) \tag{2.4}$$

Exercise physiologists have suggested that for men, A = 167 and B = −0.8. So for a 20-year-old man, the target rate would be: 167 + (−0.8*20) = 167 − 16 = 151.

For women, A = 166 and B = −0.6, so for a 20-year-old woman:

$$P = 166 + (-0.6^*20) = 154 \tag{2.5}$$

For men, each additional year of age decreases the target rate by 0.8 points. For women the decrease is 0.6 points per year.

Notice that we have different As and Bs for men and women. A and B are constant within a sex (the same for all men and for all women), but differ between sexes (different for men than for women). (We use the term sex here rather than gender on the presumption that the difference in target resting heart rate is based on biology rather than on culture.) So, A = 167 for all men and B = −0.8 for all men, but A and B are different for women. Looking at how constants vary across groups (that is, turning constants into variables) is an important way to study variation, as we will see in later chapters.

Because the equation is setting a target, there is no error term. But if we took actual heart rates for everyone in a group that was exercising, and also recorded everyone's ages, we could use the equation to predict the actual heart rate during exercise. Then we could compare the prediction for each person in the group, based on their age, to their actual resting heart rate. This would generate an E term for each person – the amount by which the prediction missed their actual heart rate. But for this example, there is no E.

E is added to the statistical equation to indicate that because of sampling error, randomization error or some other form of error we do not expect to be able to predict exactly the heart rate with age or the homicide rate with the poverty level. The prediction part of the function is A + B*X, and E is the error in prediction. We will return to the idea of separating a function for Y into a prediction and the error in the prediction in later chapters.

## Units of Analysis

In this section, we discuss the observational elements or the units of analysis that are common in social science research. Most social scientists study people. But while some of us study individuals, others study groups of people, or the organizations people form, or different spatial aggregations of people, or the results of what people do. The unit of analysis in a study is the thing on which data were actually collected. In most polls and surveys, we collect data about individual people; thus the person is the unit of analysis. But it is also common to use a survey interview with a single person to find out information about households or families or couples. Then the household or the family or the couple might be the unit of analysis.

There is a large literature examining how people interact in small groups. In these studies, researchers convene groups, let them interact, and observe what happens. In such studies researchers will collect data on a number of groups, and the group becomes the unit of analysis. Other researchers are interested in how governments work, how nations differ from one another, why some firms have progressive child care policies and others do not. For such studies the unit of analysis might be governments, nations or firms. One of the first questions you should ask about a study is, "What is the unit of analysis?"

Answering this question tells you what has been studied. It can lead to a second, critical question – is the thing being studied the same thing described by the theory being used and the same thing that the authors are drawing conclusions about? A study can only legitimately draw conclusions about the unit of analysis used in that research. Conclusions about anything else require a leap of faith. The problem of studying one kind of thing and drawing conclusions about another has a special name – the **ecological fallacy**.

The *ecological fallacy* has nothing to do with ecology as the term is used at present, but the name persists. The idea first appeared in the social science literature in a study by sociologists William Ogburn and and Inez Goltra (1919). Women had just been given the right to vote in Oregon. Ogburn and Goltra wanted to know if women would vote differently than men.

Since Ogburn and Goltra didn't know how individual women voted, they used data on the percentage of women voting in each precinct in Portland, Oregon and the percentage of people voting "no" on various propositions on the ballot. They used this data to see if there were different voting patterns in precincts with a large proportion of women voting compared to those with a small percentage. But they noted that this kind of analysis doesn't necessarily prove that women tended to vote differently than men. One cannot draw that conclusion. One can conclude that precincts with many women voters have different voting patterns, but because the unit of analysis is precincts, not people, one cannot draw conclusions about people.[5]

We can imagine that there are precincts that are very liberal and those that are very conservative. Suppose that in the liberal precincts women turn out to vote and in the conservative ones they don't. Further, imagine that in the liberal precincts most people favor some bond issues on the ballot while in the conservative precincts most people oppose the bond issues. But it may be that within a precinct, liberal or conservative, women are no more likely to favor the bond issue than men. But when we look at the data, we will see that in the precincts with lots of women voting, the bond issues are favored. Bond issues are not favored in the precincts with few women voting.

Two special units of analysis are worth mentioning. One is when we do a study of events. We might collect data on the characteristics of wars, or the characteristics of judicial decisions, or of auto thefts. In each case, the event is the unit of analysis. We may gather data about the event by looking at official records or by interviewing people, but the thing being studied is still the event. Over the last few decades the study of events has become increasingly popular in the social sciences, in part because many kinds of events are important and theoretically interesting, but also

because better and more powerful statistical tools have been developed to study events, called **event history analysis.**

The other unit of analysis that deserves special mention is time. For example, we can collect data on unemployment rates and homicide for various years for a county or a state or a nation. Then if we examine the relationship between unemployment and homicide and suicide, the unit of analysis is the year, so that we might have 25 years worth of unemployment, homicide, and suicide data for the state of Vermont.

If we are interested in studying social change, then data over time are very attractive. Such studies are called **time series analyses** (otherwise known as a **longitudinal** study). One reason they are popular is that they can help us solve the problem of **causal ordering.** Since time flows in only one direction, and no one has yet invented a practical device for time travel, we can assume that things that happened in the past are causes of, but are not caused by, things that happen later.[6] As we will see in chapter 6, being able to make assumptions about causality is very important in analyzing data.

There are some important problems with using time as a unit of analysis. Some of these are quite technical and beyond the scope of this book. But a simple and common problem is a tendency to think we see a link between two variables over time when in fact all we are seeing is a general trend that may be driven by a third variable we aren't thinking about. For example, if we plot the number of ministers in the US in the nineteenth century and the amount of rum consumed, using years as the unit of analysis, we'll find a strong positive relationship – years with lots of ministers will also be years in which lots of rum is consumed. Does this mean ministers are drinking a lot?

Not necessarily. A more reasonable explanation is that over time the population of the US grew, and as a result there were more ministers and more rum consumed, but there was no link between ministers and rum. This sort of **correlation** without a causal effect is called **spurious.** We will discuss spuriousness in more detail later. For now we simply want to note that many social variables have strong time trends in them, and those general trends can be mistaken for strong links between two variables in an analysis with time series data.

## Data Structure

There are two basic data structures and one hybrid. The basic structures are *time series,* discussed above, and **cross-sectional.** Cross-sectional studies are those in which we collect data on many examples of the unit of observation at one point in time. In a cross-sectional study we might conduct a survey of people by interviewing all of them over a period of a few weeks (which, for the purposes of the study, can be considered interviewing them all at the same time).[7] Or we might collect data from official records on crime rates and unemployment rates for cities, using many cities and gathering data for the same year from each city. In contrast a time series study takes one object (say a country or a city) and collects data over time – every month or year.

One advantage of cross-sectional studies is that it is usually possible to collect more data than can be acquired for a time series study. And if you have a new idea for something to measure, you can collect those measurements on a cross-section in a short time and do your study. But if you use a time series, you of course have to collect your measurements over time. Thus if you can only make one measurement a year, you may have to wait a long time before you have enough data to analyze. This is why most time series studies use data from official records or other historical sources. But the disadvantages of a cross-sectional study is that the time ordering of **causation** is not available to provide leverage in determining what's causing what, so theory has to bear more of a burden in making assumptions about cause and effect.

The hybrid approach is called a **panel study** or pooled time series cross-section. Unlike the time series study, we collect data over time on many units. Unlike the cross-sectional study, you observe the units at several points in time. The panel design provides a large sample size and time ordering to help with assumptions about causality.[8] For example, a very famous study, the US Panel Study of Income Dynamics started by interviewing about 5,000 families in 1968. The families were re-interviewed every year until 1997 and are now being re-interviewed every two years. The data set has information on about 5,000 families for about 30 years, or 150,000 observations!

The US Panel Study of Income Dynamics has provided extremely valuable information about the economic life of American families. We have learned about how people move into and out of poverty, what role welfare has played in getting people out of poverty, how women's work outside the house has changed and how that change has influenced housework and childbearing, and many other important issues. Many of these analyses could not be done with a cross-sectional survey because while we would know how families differ at one point in time we wouldn't know how families change over time. Nor could following one or a few families over time give us an understanding of the diverse experiences that American families have faced over the last three decades.

We can also do panel studies when the unit of analysis is a state or nation or other unit for which official statistics have been collected for a long period of time. When someone else is collecting and storing the data, it can be rather easy to put the data together into a panel. But the researcher must be careful to check that the ways data have been collected have not changed over time.

In the example of state homicide rates, we could collect from the official data sources information on homicide rates and other variables for all the states for all the years for which the data exist. This would give us a larger sample size and also allow us to look at how things have changed over time, including looking at how changing economic conditions may have influenced homicide or how changes in laws like the increased use of the death penalty have influenced homicide rates. Again, these kinds of analysis either cannot be done with a single cross-section or are much less powerful.

The example of state homicide rates can make clear the differences between panels, cross-sections, and time series. The panel data set would use all states and

all years for which data are available. The cross-sectional study (which is the design of our example) will use all states, but only one year (sometimes variables are not all from exactly the same year because of what's available when we do the study, but we treat the data as if it's all from the same point in time). A time series study would take one state and look at changes in homicide rates and other factors over time. Each design has its strengths and weaknesses, although panel data are generally the best if they can be collected with high quality. But that can be very expensive and time consuming.

## Sample Size and Sample Selection

If we do not have data on every unit, we have a sample. The **sample** is a subset of all the units on which we would like to have data, and we refer to all those units as the **population**. How many observations are there and how were individual units selected to be in the sample? As we will see later, the more observations we have the better statistical tools work and the more likely we are to find interesting and subtle results. But even more important than the number of observations is the way observations were selected.

We need to be very careful about how we select individuals to be in our sample. In some studies, especially those relying on official statistics or historical records, researchers include all observations for which the data are available. By comparing these two kinds of studies, we can get a sense of why sample selection is important. In reading a study, one of the first questions to ask is: How was the sample selected?

In surveys, a great deal of care is taken to insure that the sample is representative and generated by a **random selection process** that is well understood by researchers. Recall from chapter 1 that a random selection process gives everyone in the population an equal chance of being included in the sample. It chooses people from the population rather like drawing balls from the hopper for a lottery or throwing a die. If the lottery is fair, every number (every numbered ball) has an equal chance of being selected. If the die isn't loaded, then each side has an equal chance of coming up.

We call randomly selected samples **probability samples** because we know the probability that each person in the population will be selected for the sample. We use random selection because it is representative and because we know how random samples behave and can use that understanding to make careful statements about the data and the population from which it was drawn. In other words, if we have data from a probability sample, we know how sampling error behaves and can account for it in our models.

When a historical record-keeping process generates the sample, the sample is drawn from a population of all units that might have such data. But it is not usually a random sample in which every unit has the same probability of being in the sample. This can introduce biases into our analysis. For example, if we use data on national emissions of pollutants, we find that mostly the richer countries gather and report

such data. Thus our sample is mostly of rich countries and our conclusions must take that into account.

Some years ago there was an influential study of birth rates that used data from all countries from which data were available. It was a number of years before a critical reader checked the original data and found that the entire sample for Asia consisted of only two countries, Israel and Japan. Only Israel and Japan had reported data on the key variables. It is hard to draw valid conclusions about fertility patterns for the world if only two countries are used to represent the whole of Asia. No sample using historical records is perfect, just as no survey is perfect. We can learn a lot from imperfect samples if and only if we are careful to understand the flaws in the sampling process.[9]

## Box 2.4 How Representative Are Surveys?

The US General Social Survey (GSS), one of the surveys participating in the ISSP (data used for Example 2 throughout the text), has a large proportion of women. The Census Bureau, which has the most accurate data on the US population, tells us that the proportion of women in the US population is 51.2 percent. The GSS is considered one of the higher quality data sets used on a routine basis in the social sciences. But in all surveys there are limitations. One of the limits of the methods used in the GSS is that there tend to be a higher proportion of female respondents and a lower proportion of male respondents than would be true if we interviewed everyone in the population. In the 2000 ISSP data set, 56.1 percent of American respondents are female.

Many government surveys that are used to make policy take special steps to try to insure that the sample is representative. There's an old saying "Close enough for government work," indicating that work for the government can be sloppy. But in the world of surveys the US government produces some of the highest quality data in the world. In this case "Close enough for government work" means the most precise work in the world. In contrast, many public opinion polls used by the news media and many marketing research surveys don't take much trouble to insure a representative sample. So it might be more accurate to say, "Close enough for the private sector."

Of course, there are exceptions to every rule. Some public opinion polls and marketing studies are done very carefully. When you are presented with results from a survey or other study, you should ask yourself: "Was this research done carefully? Are the results good enough for the purposes they are being used?" If the purpose is to get some general sense of public feeling to satisfy curiosity, not much precision is needed. If major decisions hinge on the data analysis, such as the allocation of federal funds to areas with high levels of poverty or unemployment, great care must be taken. That's why the federal government does such a good job at its statistical analyses – the results matter.

## Box 2.5 Large and Small Samples

We have more confidence in results from larger samples than from smaller samples if both the large and small samples are random samples. But the quality of the process by which the data were collected is more important than the size of the sample. Here is one famous example. In 1936 the *Literary Digest*, a popular US magazine of the time, conducted a very large survey to predict the outcome of the US Presidential election in which Republican Alf Landon was running against incumbent Franklin Delano Roosevelt (Bryson, 1976). The *Literary Digest* sent questionnaires to 10 million people and got 2.3 million responses, a large sample by any standard. But they miscalled the election, predicting that Landon would win. At the same time, a young social statistician, George Gallup, did a survey of 10,000 people and not only called the election correctly but predicted the *Literary Digest* poll would be wrong. How did this happen?

The *Literary Digest* had plenty of data. But they drew their sample from lists of magazine subscribers, car owners and phone books, as well as a few lists of registered voters. In 1936, the world was in the midst of a depression, and so the *Literary Digest* lists were biased towards the wealthy. And by having such a low response rate (23%) they were probably getting people most concerned about the election and thus those most dissatisfied with Roosevelt. In contrast, Gallup used a sample similar to the random samples we use now, and got the right answers. How you draw the sample is more important than sample size.

## What Have We Learned?

In this chapter we have explained the terms used in quantitative data analysis. In doing so, we have begun to introduce the concepts that are used in statistics. We have variables that take on different values for different observations. They may be at any of three levels of measurement (nominal, ordinal, or interval), and we have to be careful about what kinds of techniques we use, depending on what kinds of measurements we have.

We also have constants – numbers that are the same for everyone in the data set. Functions link independent and dependent variables using constants to show the relationship between the two. We can study a variety of different units of analysis, including people, groups, institutions, and events. And we can collect data from many units at one point in time, one unit at many points in time, or many units at many points in time.

But whichever data structure we have, we must pay attention to the processes by which our units of analysis ended up in the data set we are studying. Did we have a random sampling method? Did we use all of the data available? Given these various aspects of data, it is now time to begin to look at analysis tools that help us make sense of the data, starting with graphic displays.

## Applications

In this Applications section we will look at the definitions, sources for, and properties of some of the key variables we use in each of our continuing examples. We will first define the key dependent variable in each of the examples, and we also describe one of the independent variables. We'll define other independent variables as we use them in later chapters. A description of all the variables we use and the data sets that contain them will be included in Appendix 1.

### **Example 1:** Why do some US states have higher than average rates of homicide?

The model we examined in chapter 1 was:

$$\text{homicide rate} = f(\text{poverty rate}) + E \tag{2.6}$$

Our idea was that the homicide rate might depend on the poverty rate. In the Applications for chapter 1 we also mentioned a number of other theories of what might cause homicide rates to vary from state to state.

*Dependent variable: homicide rates*

In most research that examines this topic, homicide rates are based on secondary data, data that were compiled for purposes other than our research. For example, in the Baron and Straus (1988) study, they explain that the homicide rates they used were collected by the Uniform Crime Reports and consist of the 1980 rate per 100,000 population of homicides known to police. In our example, we'll use the data from 2003 provided by the US Census Bureau. State data are measured in homicides per 100,000 population.

*Independent variable: poverty*

In chapter 1 and throughout the book we will look at the poverty rate as a possible cause of homicide rates. The US Census Bureau computes the poverty line using income "thresholds," which are cut-off income values for how much income is needed annually to afford the basics of day-to-day life. Rather than identifying one poverty line for everyone in the United States, there are several poverty lines, which vary by composition and size of families. In 2002, for instance, poverty was defined at an annual income of $14,393 or less for a three-person household with no children and at $22,509 or less for a five-person household with two children. Data on percent of the population in poverty for each state were taken from the US Census Bureau for 2002.

*Labeling variables*

Homicide rate is labeled homicide03, and poverty rate is labeled poverty02.

*Units of analysis*

The units of analysis are the 50 US states. The conclusions drawn from the study thus refer to US states and not individuals.

*Levels of measurement*

The dependent and independent variables, homicide and poverty rates, are measured at the interval level.

*Data structure*

The data structure is a cross-section. Most of the data are from the years 2000 to 2003 from the US Census Bureau.

## Advanced Topic 2.1 Panel versus Cross Section for the State Data

Since most of the variables we might use are available for most years, it would be possible to construct a panel data set of all states and all years from, say 1980 to 2000. This would increase the sample size from 50 to 1,000 (20 years times 50 states) and would allow for some analyses that aren't possible with the cross-section for a single year. If we wanted to draw strong conclusions about homicide constructing the panel would be a good idea. But for purposes of understanding statistics, the cross-section is much simpler to use.

*Sample selection*
The entire population is in the study: all 50 US states. There is no need to estimate population parameters because we know the value of those parameters from the data collected.

**Example 2:** Why do people differ in their concern for animals?

Again, we propose a model:

$$\text{animal concern} = f(\text{gender}) + E \tag{2.7}$$

This model suggests that knowing an individual's gender will help us predict one's attitude towards animals. But the error term suggests that we don't expect the prediction will be perfect. To test this model, we need data. One of largest compilations of international data on the attitudes of residents in different countries in the world is the International Social Survey Programme (ISSP) (www.issp.org). Since 1985 many countries have conducted cross-sectional surveys with samples of their residents focusing on a mutually agreed upon research topic, using the same questions (after translation) in all participating countries. Some countries conduct interviews annually, others every other year, and some less frequently, so the countries participating in a given year vary. The focus of the survey changes from year to year, and has included environment, family and gender roles, religion, and social relationships. In 2000, the survey topic was attitudes on the environment and also included questions about basic demographics, including gender, ethnicity and income. One of the questions measured concern for animals. Twenty-six countries administered the environmental survey in 2000.

*Dependent variable: animal concern*
We have one question that taps respondents' views about animals:

> It is right to use animals for medical testing if it might save human lives.

Respondents were asked if they "strongly agreed, agreed, neither agreed nor disagreed, disagreed or strongly disagreed" with each statement. People who have strong concerns about animals will tend to disagree with this statement.

## Advanced Topic 2.2 Multiple Measures

It is good practice in a survey to ask multiple questions on the same topic and to word the questions so that a person with strong views on the subject sometimes has to say "agree" and sometimes "disagree." The logic behind this is that if someone is not thinking much when answering the questions and is just saying "agree" to every question, she or he will seem to be pro-animal on one question and anti-animal on another other. When we combine

the items in the scale, such a person will be in the middle rather than at either extreme. Unfortunately in this data set, we only have one survey question that measures animal concern. In constructing surveys there is always a tradeoff. Asking a lot of questions on the same topic allows for better measurement (and less measurement error) but increases respondent fatigue and the proportion of potential respondents who don't complete the survey (which increases sampling error).

If we were designing a study focused on concern with animals we would have asked at least a few questions, not just one. But in making their tradeoffs about the content of the survey the designers of the ISSP decided to go with just one question. This is often a problem with using "secondary data," – data that were collected by someone else. The researchers who designed the study may not have had the same interests that we do and may not include as many measures of what we are studying as we would like. But the advantage of such secondary data in this case is that we have data from a large international sample of high quality that would be very expensive for us to collect on our own.

We want our measure to have high scores for people who have pro-animal attitudes and low scores for those that have anti-animal attitudes. Individuals who disagreed with the question are more pro-animal than individuals who agreed. A person who "strongly agreed" has a value of 1 on this question, 2 for "agreed," 3 for "neither agreed nor disagreed," 4 for "disagreed," and 5 for "strongly disagreed," so we do not have to change the scoring.

*Independent variable: gender*

This is a variable that is almost always included in a survey. We call variables such as gender, age, and education "demographics." The interviewer recorded the "sex" of the respondent based on their judgment from the interview. So in a sense this is the interviewer's interpretation of the respondent's gender. The variable has two attributes: female and male.

*Labeling variables*

The dependent variable might be labeled "animalx" for animal concern index, and the independent variable could of course be labeled simply "sex."[10]

*Units of analysis*

The units of analysis in this problem are individuals, and the data are collected from individuals via surveys.

*Levels of measurement*

The animal concern question is measured at the ordinal level. Throughout the book, we often treat the item as if it is measured at the interval level, so that we can use this example with statistical procedures that require interval measurement.

Gender is a nominal level variable – the attributes of gender are labels or categories with no numerical value attached to them (female, male). Remember that

we often assign a number to the categories of gender, i.e., male = 1, female = 2, but these numbers have no numerical meaning; they are only labels. (However, recall that when we define binary nominal variables with the values 0 and 1, we can do some kinds of arithmetic, liking taking the average, which is then just the proportion of people in the category labeled 1.)

*Data structure*
The data in the ISSP are cross-sectional.

## Advanced Topic 2.3 Panel versus Repeated Surveys

The ISSP has compiled surveys every year since 1985. But the countries interview a different sample of people each time, so it is not a panel study. A panel study, like the US Panel Study of Income Dynamics, interviews the same people over and over again (see http://psidonline.isr.umich.edu/ for more details on this study). The panel study allows us to follow changes in an individual and within a family over time. But panel studies are much more complicated and expensive to conduct then a series of annual cross-sectional studies.

*Sample selection*
The 2000 ISSP survey consists of samples of residents in 26 countries: Austria, Bulgaria, Canada, Chile, Czech Republic, Denmark, Finland, Germany, Great Britain, Ireland, Israel, Japan, Latvia, Mexico, Netherlands, New Zealand, Northern Ireland, Norway, Philippines, Portugal, Russia, Slovenia, Spain, Sweden, Switzerland, and the United States. Each country translates the survey into its primary language and administers the survey. Sample sizes range from 745 residents in Northern Ireland to 1,705 residents in Russia. The combined sample size for all 26 countries is 31,042.[11]

Because we are using data from all the countries in the 2000 ISSP, we have to restrict our choice of variables to those that can be used meaningfully across such a broad range of countries. Gender is of course socially constructed so gender socialization, gender roles, the structural influences that impinge on men's and women's lives, and other gender-related factors will certainly differ across the countries in the data set. This will be true of other independent variables, such as marital status, education, and age that we will use to explain concern with animals in other chapters. After looking at the analyses that combine data from all countries, it would make sense in a research project to look at the data separately for each country and to explore how the effects of gender and other independent variables differ across countries and why those differences exist. Sophisticated methods for doing such analysis are the subject of a lot of research in the last few years (Byrk and Raudenbusch, 1992; Steenbergen and Jones, 2002; Western, 1998). These methods are beyond the scope of our presentation in this book. But in exercises for your class, it would be possible to pick one or a few countries and replicate the analyses we are doing for all countries to see how the models for animal concern might vary across countries.

In the rest of the book we will usually use the ISSP data on animal concern from all countries as a single data set rather than breaking it out by countries. This will give you a sense of how statistical tool work with a very large survey data set. But note that when we use the data for all countries at once we do not have a representative set of countries of the world, just the ones where social science researchers were interested in participating in the ISSP and were able to find the resources to do so. So we can't generalize to the world from the data set. And since the sample size for each country is not proportional to the populations of the country, the sample that uses data from all countries is not directly representative of the whole population of those 26 countries. So when we use the data for all countries with taking direct account of the individual countries in the analysis, we are doing something that is useful as an exercise in working with a large sample, but we would not do those kinds of "all countries" analyses for research. Rather, we would take explicit account of the countries to compare them. But you can replicate our "all countries" analyses with data for a single country and get results that are perfectly sound for that country. In the Applications at the end of Chapter 14 we do examine the effects on animal concerns of being in a particular country.

## Advanced Topic 2.4 Sampling Approaches in the ISSP

The details of how each country generated their samples are quite complicated and vary by country. Some countries selected representative samples of the entire population, while others covered only major cities or are restricted in other ways. Some countries carried out face-to-face interviews, while others used mailed or self-administered surveys. The details on the sampling and data collection procedures are described in detail on the ISSP website at (http://www.gesis.org/en/data_service/issp/data/2000_Environment_II.htm).

### **Example 3:** Why are some countries more likely to ratify environmental treaties than others?

Here is our model:

$$\text{environmental treaty participation} = f(\text{voice and accountability}) + E \qquad (2.8)$$

One reason countries may differ in the extent they participate in environmental treaties is because they differ in how much a nation's citizens have "voice" and the extent to which a government is accountable to its citizens. But we don't expect voice and accountability to perfectly predict environmental treaty participation, so we also include an error term in the model.

*Dependent variable: environmental treaty participation*

Between 1946 and April 1999, there have been 22 multilateral international environmental treaties. Treaty topics include air pollution, oil pollution, and greenhouse

gases. The idea of looking at how many treaties a country has ratified was introduced by Dietz and Kalof (1992). In a more extensive analysis of this issue, the total number of treaties a country participated in was computed for 191 nations by Timmons Roberts, Parks and Velasquez (2004; see the article for a complete description of all the treaties and the sources of information the authors used to determine treaty participation).

*Independent variable: voice and accountability*

It may be that governments that are unaccountable and oppressive will be more likely to ignore the demands of environmentalists and the international community to act in environmentally responsible ways. To measure this, Kaufmann, Kraay, and Zoido-Lobaton (2002) created an index called "Voice and Accountability." This index includes numerous measures of citizens' political rights, civil liberties, aspects of the political process, and how independent the media are. Scores on the scale reflect the extent citizens of a country are able to participate in the selection of their government officials and have freedom of expression. The inclusion of how independent the media are serves to measure how closely authority figures are monitored and held publicly accountable for their behaviors. The index is created by several measures that come from various non-governmental organizations (NGOs), risk rating agencies, and think tanks. Scores can range from −2.5 to +2.5. While the computation of the index is quite complicated, what is important to note is that higher scores reflect greater levels of "voice and accountability," meaning more freedom of expression, a free press, and considerable citizen participation in government. Lower scores, on the other hand, reflect little voice and accountability.

*Labeling variables*

Typing (or saying) number of environmental treaties and voice and accountability rate is rather cumbersome. We could call the former X and the latter Y, but if we have lots of variables it can be hard to remember what is what. As we noted in chapter 1, sometimes it's convenient to use short strings of letters as names for variables. When we created this data set, we told the computer to call number of environmental treaties "envtreat" and the voice and accountability index "voice." This helps us remember what is what.

*Units of analysis*

Our units of analysis are nation states as recognized by the World Bank and other sources for our data. Our theory and any conclusions we draw should thus be about nations, not about individuals, communities or any other unit.

*Levels of measurement*

Both variables are interval level variables. Environmental treaty participation is measured from 0 to 22. The voice and accountability scale is scored from −1.93 to +1.73.

*Data structure*
Treaty participation is measured over six decades, so it might be thought of as a time series while voice and accountability is measured at one point in time. However, we are not looking at how participation in treaties changed over time but rather a summary measure of how many treaties a country had ratified by 1999.

*Sample selection*
We were able to get data on environmental treaty participation for 191 countries. Data were available on the voice and accountability scale for 169 of these countries. This includes many small countries; in fact 22 percent of the countries (N = 42) have populations of less than 1 million. Sometimes in analysis of cross-national data analysis researchers restrict their sample to large nations, for example only considering nations with populations of over 1 million. For some kinds of analyses it doesn't make sense to compare countries like China, the US, and the UK with very small countries like Trinidad and Tobago or Fiji. Since participating in environmental treaties is a political activity, and both large and small countries have political processes, we don't think the size restriction makes sense here. So we will use all countries for which we have data. Some researchers have theorized that the size of a political unit has a substantial effect on its political processes (Dahl and Tufte, 1973; Dietz, 1996/1997; Frey and Al-Mansour, 1995) so we could look at population size, population density or rate of population growth as predictors of treaty participation. We don't pursue those here but they might be interesting topics for further research.

## **Example 4:** Why do people differ in their knowledge of how the AIDS virus is transmitted?

Our model is:

$$\text{AIDS knowledge} = f(\text{gender}) + E \tag{2.9}$$

This model suggests that knowing an individual's gender will help us predict his or her knowledge about AIDS transmission. But the error term suggests that we don't expect the prediction will be perfect. To test this model, we will use data from the 2000 Ugandan DHS dataset (see http://www.measuredhs.com/pubs/pdf/FR128/01Chapter1.pdf for a summary of the entire 2000 Ugandan-DHS survey, including background information on Uganda, how the sample was collected, and a summary of the survey data findings).

*Dependent variable: AIDS knowledge*
The interviewer asked each respondent to state spontaneously what they think can be done to prevent the transmission of the AIDS virus (common answers, for example, were abstinence from sex, having only one sexual partner, not engaging in prostitution). After respondents gave their complete list, they were also probed about various ways AIDS can or cannot be transmitted. We will focus on the

transmission of AIDS and condom use and whether respondents knew that condom use can help prevent transmission of the disease.

The response categories were "yes," "no" and "don't know." Unless otherwise indicated, throughout this book, "no" and "don't know" responses are combined since both responses reflect a lack of knowledge that condoms can be used to prevent the transmission of AIDS.

*Independent variable: gender*

Again, this is one of those basic demographic variables that is almost always included in a data collection effort. The variable has two attributes: female and male.

*Labeling variables*

The dependent variable might be labeled "AIDScon" for knowledge that condoms can reduce the likelihood of transmitting AIDS, and the independent variable could of course be labeled simply "sex."[12]

*Units of analysis*

The units of analysis in this example are individuals, and the data are collected from individuals via face-to-face interviews.

*Levels of measurement*

If we combine "don't know" and "no" responses, we have a nominal level variable that could be scored 1 for those who said yes (i.e., knowledgeable about condom use and AIDS transmission) and 0 for those who said no/don't know (i.e., lacked knowledge about how to prevent AIDS transmission). We can calculate the mean (the proportion saying yes), which is 0.77 or 77 percent. That is, about three-fourths of respondents know that condom use can reduce the likelihood of transmitting AIDS.

Gender is also a nominal level variable – the attributes of gender are labels or categories with no numerical value attached to them (female, male).

*Data structure*

The data in the Ugandan-DHS survey are cross-sectional.

*Sample selection*

A stratified, clustered sampling design by region of the country was used to select the Ugandan-DHS sample. This is not a representative sample of Ugandan citizens though. Conducting a large-scale survey in a country like Uganda is quite challenging. First, some parts of the country were deemed too unsafe for interviewers to travel. Since most residents live in rural areas, urban area residents were oversampled. In addition, over three-fourths of the sample was female. Since other goals of the survey were to assess women's fertility behaviors and maternal and children's health, there was a particular interest in surveying women. The goal was to interview at least 6,500 women and 1,800 men.

## Exercises

The purpose of these exercises is not only to practice the material in the chapter but also to stimulate thinking and discussion. We present the kind of information that often comes with codebooks describing data sets or brief summaries of research papers and reports. For each of the exercises, thinking through and discussing why you answered the way you did is as important as the answer itself, because this kind of thinking hones your skill at quantitative reasoning. In some cases, the information provided with data or research reports lends itself to more than one interpretation. Understanding the basis for multiple interpretations can provide the basis for deeper understanding of the concepts.

---

1. In the 2000 International General Social Survey that provides the data for our analysis of attitudes towards animals, there are a number of variables we might use as independent variables in a model predicting attitudes towards animals. For each of the following variables, indicate whether it is nominal, ordinal, or interval. Explain your reasoning. The word in parentheses is the name of the variable in the ISSP data set. The following information provided with the data set explains what each variable represents. In every case we have left out the codes for missing data (people who didn't answer the question).

A Marital Status (marital)
"Are you currently – married, widowed, divorced, separated, or have you never been married?"
RANGE: 1 to 5:
1 married;
2 widowed;
3 divorced;
4 separate;
5 never married

B Age (age)
Respondent's age
RANGE: 18 to 99

C Education (educ)
What is the highest grade in elementary school or high school that you finished and got credit for?
RANGE: 0 to 99

D Highest degree earned (degree)
Highest educational degree earned by respondent
RANGE: 1–7:
1 None; still in school;
2 Incomplete primary school;
3 Primary school completed;
4 Incomplete secondary school;
5 Secondary school completed;
6 Semi-higher; incomplete university;
7 University completed

E Family income at age 16 (INCOM16)
Thinking about the time when you were 16 years old, compared with American families in general then, would you say your family income was – far below average, below average, average, above average, or far above average?
RANGE: 1 to 5:
1 Far below average;
2 Below average;
3 Average;
4 Above average;
5 Far above average

F Where you live (urbrural)
Describe the place where you live
Range: 1–5:
1 Urban;
2 Suburb;
3 Small city or town;
4 Country village;
5 Farm, home in the country

---

2. Suppose we think the following model might explain homicide rates:

$$Y = A + BW + E \quad (2.10)$$

Where Y is the state homicide rate, W is the percent urban and E is the error term. A good estimate of A and B for the US Census Bureau state data is A = 2.86 and B = 0.026. Table 2E.1 gives values for urbanization level and homicide rate for several states. Calculate the predicted value for homicide rate and the error value for each of these states.

**Table 2E.1** Homicide rate, percent urbanization, and predicted homicide rate for five states

| *State* | *Homicide rate* | *Urbanization* | *Predicted homicide rate* | *Error* |
|---|---|---|---|---|
| Georgia | 7.6 | 71.6 | | |
| Maryland | 9.5 | 86.1 | | |
| New Jersey | 4.7 | 94.4 | | |
| South Carolina | 7.2 | 60.5 | | |
| Wyoming | 2.8 | 65.1 | | |

3. For each of the following studies, identify the unit of analysis, the data structure, the independent variable(s), the dependent variable(s), and the levels of measurement of the variables.

A McLaughlin, P. and Khawaja, M. 2000. The organization dynamics of the US environmental movement: Legitimation, resource mobilization, and political opportunity. *Rural Sociology* 65, 422–39.

This study models the number of environmental organizations founded in the United States each year from 1895 to 1994. The most important independent variable in the study is the number of books published that year on environmental topics, which the authors interpret as a measure of the ideological climate in a given year.

B Gamson, W. 1975. *The Strategy of Social Protest.* New York: Dorsey.

This study draws a simple random sample of protest groups active in the US over a 75-year period. The dependent variable is whether or not the group was successful in achieving its goals. One of the key independent variables was whether or not the group used violent means.

C Kalof, L. 2000. Vulnerability to sexual coercion among college women. *Gender Issues* 18(4), 47–58.

This study surveyed a random sample of undergraduate women at a university in up-state New York. In that survey the author measured women's gender attitudes and experiences with sexual coercion. Two years later she sent a follow-up survey to the women who were still enrolled at the university. She again examined gender attitudes and a new variable, experiences with sexual coercion since the first survey. She found that: 1) initial attitudes (as measured by the first survey) had no effect on experiences with sexual coercion over the 2-year period; 2) initial experiences with sexual coercion did not make women vulnerable to more sexual coercion over the 2-year period; and 3) attitudes were not changed by experiences with sexual coercion over the 2-year period.

D St Lawrence, J. S. and Joyner, D. J. 1991. The effects of sexually violent rock music on males' acceptance of violence against women. *Psychology of Women Quarterly* 15, 49–63.

This study examined the effects of rock music on males' attitudes toward women. Undergraduate men were randomly assigned to listen to one of three types of music: sexually violent heavy-metal rock, Christian heavy-metal rock, or easy-listening classical. Participants were administered the Attitudes Toward Women Scale before and after exposure to the music. The authors found that exposure to either kind of heavy-metal rock music, regardless of lyrics about sexual violence, increased males' sex role stereotyping and negative attitudes toward women.

4. Here are some variables that have been used in the US General Social Survey and ISSP data sets. In exercises in later chapters, we will be analyzing many of these variables. The variable name is listed first, followed by a definition of the variable and how it is coded. Identify the level of measurement of each of these variables.

a) Childs = Number of children an individual has.
b) Worklife = "How successful are you in your work life?"
   1 not at all successful;
   2 not very successful;
   3 somewhat successful;
   4 very successful; and
   5 completely successful
c) Hrs2 = Number of hours usually worked in a week.
d) Wrkslf = Are you self-employed or do you work for somebody else?
   1 self-employed; and
   2 work for somebody else
e) Spwrksta = Spouse's work status:
   1 working full-time;
   2 working part-time;
   3 temporarily not working;
   4 unemployed/laid off;
   5 retired;
   6 in school;
   7 keeping house; and
   8 other.

---

5. In the previous chapter, we learned about independent and dependent variables. For each of the following hypothetical statements, identify the independent and dependent variables.

a) Canadians whose first language is English had higher average scores on a scale of patriotism compared to those Canadians whose first language is French.
b) The more hours spent each week in leisure activities, the higher a person's life satisfaction.
c) Higher infant mortality rates tend to be concentrated in countries that are less technologically advanced.
d) Children who played a video game that had violent content were more likely than children who were exposed to a non-violent video game to act aggressively in the two weeks following video game exposure.
e) The conservative country of Italy has had a much lower rate of cohabitation before marriage than the more liberal countries of Canada and the US.

## References


Baron, L. and Straus, M. A. 1988. Cultural and economic sources of homicide in the United States. *The Sociological Quarterly* 29, 371–90.

Binder, A. 1984. Restrictions on statistics imposed by method of measurement – some reality, much mythology. *Journal of Criminal Justice* 12, 467–81.

Bollen, K. A. and Barb, K. H. 1981. Pearson's r and coarsely categorized measures. *American Sociological Review* 46, 232–9.

Bollen, K. A. and Barb, K. H. 1983. Collapsing variables and validity coefficients – reply. *American Sociological Review* 48, 286–7.

Bryson, M. C. 1976. The *Literary Digest* poll: Making of a statistical myth. *The American Statistician* 30, 184–5.

Byrk, A. S. and Raudenbusch, S. W. 1992. *Hiearchical Linear Models.* Newbury Park, CA: Sage.

Dahl, R. A. and Tufte, E. R. 1973. *Size and Democracy.* Stanford, CA: Stanford University Press.

Demographic and Health Surveys (DHS). 2000. *Ugandan-DHS 2000 Survey.* www.measuredhs.com. Calverton, MD.

Dietz, T. 1996/1997. The human ecology of population and environment: From Utopia to Topia. *Human Ecology Review* 3, 168–71.

Dietz, T. and Kalof, L. 1992. Environmentalism among nation states. *Social Indicators Research* 26, 353–66.

Ferrando, P. J. 1999. Likert scaling using continuous, censored, and graded response models: Effects on criterion-related validity. *Applied Psychological Measurement* 23, 161–75.

Frey, R. S. and Al-Mansour, I. 1995. The effects of development, dependence and population pressure on democracy: The cross-national evidence. *Sociological Spectrum* 15, 181–208.

Gaither, C. C. and Cavazos-Gaither, A. E. 1996. *Statistically Speaking: A Dictionary of Quotations.* Bristol, England: Institute of Physics Publishing.

Hattam, V. 2005. Ethnicity & the boundaries of race: Rereading Directive 15. *Daedalus* 134, 61–9.

Hollinger, D. 2005. The one drop rule & the one hate rule. *Daedalus* 134, 18–29.

International Social Survey Programme (ISSP). 2000. 2000 Environment II data set. www.issp.org. Catalog no. ZA 3440. Cologne, Germany: GESIS-ZA Central Archive for Empirical Research.

Kaufmann, D., Kraay, A., and Zoido-Lobaton, P. (Jan. 2002). *Governance Matters II: Updated Indicators for 2000/01.* Policy Research Working Paper no. 2772. The World Bank Research Development Group and World Bank Institute; Governance, Regulation and Finance Division (http://hdr.undp.org/reports/global/2002/en/).

King, G. 1997. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data.* Princeton, NJ: Princeton University Press.

Krieg, E. F. 1999. Biases induced by coarse measurement scales. *Educational and Psychological Measurement* 59, 749–66.

O'Brien, R. M. 1983. Rank order versus rank category measures of continuous variables – comment. *American Sociological Review* 48, 284–6.

Ogburn, W. F. and Goltra, I. 1919. How women vote: A study of an election in Portland, Oregon. *Political Science Quarterly* 34, 413–33.

Prewitt, K. 2005. Racial classification in America: Where do we go from here? *Daedalus* 134, 5–17.

Roberts, J. T., Parks, B. C., and Vasquez, A. A. 2004. Who ratifies environmental treaties and why? Institutionalism, structuralism and participation of 192 nations in 22 treaties. Global Environmental Politics 4(3), 22–64.

Snipp, C. M. 2003. Racial measurement in the American census: Past practices and implications for the future. *Annual Review of Sociology* 29, 563–88.

Steenbergen, M. R. and Jones, B. S. 2002. Modeling multilevel data structures. *American Journal of Political Science* 46, 218–37.

Stevens, S. S. 1946. On the theory of scales of measurement. *Science* 103, 677–80.

Uganda Demographic and Health Surveys. 2001. Calverton, Maryland: UBOS and ORC Macro (http://www.measuredhs.com/pubs/pdf/FR128/00FrontMatter.pdf).

US Census Bureau. 2000. Table 33. Urban and rural population, and by state: 1990 and 2000 (http://www.census.gov/prod/cen2000/index.html).

US Census Bureau. 2002. Historical poverty tables: Table 21. Number of poor and poverty rate, by state: 1980 to 2006. Year 2002 (http://www.census.gov/hhes/www/poverty/histpov/hstpov21.html).

US Census Bureau. 2003. Table 295. Crime rates by state, 2002 and 2003, and by type, 2003 (http://www.census.gov/prod/2005pubs/06stata; www.census.gov/prod/2005pubs/06statab/law.pdf).

Velleman, P. F. and Wilkinson, L. 1993. Nominal, ordinal, interval and ratio typologies are misleading. *The American Statistician* 47(1), 65–72.

Western, B. 1998. Causal heterogeneity in comparative research: A Bayesian hierarchical modeling approach. *American Journal of Political Science* 42, 1233–59.