CHAPTER 1
AN INTRODUCTION TO QUANTITATIVE ANALYSIS

Outline

Why do some states have high homicide rates while in other states the occurrence of a homicide is very rare? Why are some countries more likely than others to participate in environmental treaties? Why do some people feel animals have rights while others feel animals can be treated as objects? Why do some people know how the AIDS virus is transmitted and others don't?

In this text we will explore each of these questions using quantitative methods. We will try to answer them by developing models that help us understand why people, states within the US, or countries in the world differ from one another. In attempting to answer these questions we will introduce the standard tools of modern quantitative analysis in the social sciences – statistics. Our answers will always be tentative and never certain. But the scientific method applied by using statistical tools can help us make better decisions about how the world works.

Of course, this is a statistics text, not a book about homicide or the environment. The questions we examine are intended to introduce the tools of statistics. Once you understand the tools, you will be able to see how they can be used to answer many other questions across a range of issues. Perhaps more important, they can help you to think critically about research that is presented to persuade – whether in a paper from a scientific journal, a technical report from a government agency, or in a newspaper story.

Each question we pose in the first paragraph is about variation – why do some people, states or countries differ from others? We attempt to understand that variation by building models. In quantitative analysis we use the term model in much the same way as it is used in everyday life. A model is a representation of something. For example, a model train captures the look of a real train, but in miniature, and a model apartment shows you what the apartment you are thinking of renting might be like. A fashion model shows how you might look wearing the clothes being modeled.[1]

In quantitative analysis we build models with numbers. We want to understand why the things being studied vary – why people are different from one another, why states differ, why countries differ. Like the models of everyday life, quantitative models are useful if they capture the key features of what we are studying. But they also simplify reality and can be deceptive if we don't look at them with a critical eye.

Building models and using them to understand variation is a central theme in quantitative analysis. In this text, we explain how models are developed, show how they can be used to explain variation, and examine how models of variation relate to the ongoing dialogue that is science.

Statistics is intended, in large part, to deal with models that include some error. All of our models will be imperfect descriptions of reality, and the difference between the model and what we observe in the real world is the error in the model. This idea of error is a bit removed from the everyday sense of model trains, model apartments and fashion models. But there is an analogy. As makers of model trains strive to add more and more detail, they are in a sense trying to reduce the error – the difference between the model and reality. If the actual apartment you rent does not

much resemble the model you were shown, you have a sense that things are wrong. And taking fashion models with very unusual physiques as a representation of the typical man or woman is clearly an error.

Since all data include error, we will discuss the kinds of error that are most important in social science data. This is the starting point for understanding how statistics allows us to take error into account in our models. As we will see, the error is a critically important part of our models, and we will think as hard about the error as we do about the rest of the model.

What is Statistics?

What does the term statistics mean? There are two definitions of statistics in the typical dictionary (Brown, 1993):

1 the field of study that involves the collection and analysis of numerical facts or data of any kind; and
2 numerical facts or data collected and analyzed.

The first definition refers to the field of study to which this book provides an introduction. The second definition refers to what, in some sense, statisticians study – numerical (or quantitative) data. This is the everyday use of the term statistics – the numbers that are intended to represent some aspect of life. Everyone encounters sports statistics, statistics on cars, statistics on how the economy is doing and so on.

> *A third and more technical definition of statistics* is discussed at the end of this chapter as an Advanced Topic.

The use of such quantitative data goes back at least to the earliest city-states. We know the Babylonians and Egyptians collected numerical data on crops, for example. In fact the term statistics has its roots in the Latin word for "state" indicating the historical linkage between the government and numerical data.

As a field of study, statistics develops tools that allow us to generate better numbers to describe the world. As we noted above, all models of the world involve some error. One of the major concerns in the field of statistics is to understand how error may enter our models and in the numbers that we use to describe the world. By understanding these errors we may be able to reduce them substantially. And even when we can't make them small enough to ignore, statisticians have given us tools to help us understand how large the errors may be. This allows us to guard against making decisions that treat what may be error as if it were fact.

Models to Explain Variation

To understand and use quantitative methods, we must have some sense of the process of proposing models, criticizing them, and learning from the process. Generally, a model has the form:

$$Y = f(X) + E \tag{1.1}$$

We say this as "Y equals f of X plus E."

Sometimes the equation is shown with a small subscript i after Y, X, and E. Then the equation would be $Y_i = f(X_i) + E_i$. This is sometimes done to emphasize that every observation – every person in a survey, every country in a cross-national study – can have its own value for Y, X, and E. Of course, two observations may have the same scores on a variable. Two people might have the same level of education or income or two states might have the same homicide rate. But X, Y, and E can vary from person to person even if not every person has a unique value on each of the variables. We won't use the subscripts because they can be confusing when you are first learning statistics.

X and Y can vary across observations, so they are called **variables**. Y is the **dependent variable**, the thing we are trying to explain. If we are trying to understand why states might vary in their homicide rates, then Y would be the homicide rate. If we are trying to understand who knows that the transmission of AIDS is reduced by condom use and who doesn't, Y would be each person's response to a survey question about AIDS transmission.

X is the **independent variable**. The equation suggests that the variation in Y across observations may be the result of variation in X. The equation implies that because X is different from observation to observation then Y will be different. In explaining homicide rates we might think that homicide is a result of poverty, so X would be some measure of poverty in a state. For example, we could use the percent of the state's population below the federal poverty line for X. We might think that a person's knowledge about the AIDS virus depends on their gender, so then X would be gender. One helpful way of remembering the difference between dependent and independent variables is that the dependent variable *depends* on changes in the independent variable.

In some fields, terms other than independent and dependent variables are used to describe the thing we are trying to explain and the thing used to explain it. For example, for obvious reasons the dependent variable is sometimes referred to as the "left hand side variable" and the independent variable is called the "right hand side variable." And sometimes the dependent variable is called the response (because it is responding to the independent variable) while the independent variable is called the "carrier," the "predictor variable," or the "covariate."

The "f" in the equation is just an abbreviation that covers any way X might be linked to Y. You may remember "f" from algebra. It means "is a **function** of." It says that there is a relationship between X and Y, but it doesn't say what that

relationship will look like. It might be that as X gets larger Y gets larger. That is what our theory of homicide is saying: states with more poverty will have higher homicide rates. But the "f" by itself is not that specific. It could allow for Y to get smaller as X gets larger, or for some more complicated relationship. For example, using just the "f" allows for the possibility that the states with the highest and lowest poverty rates have the lowest homicide rates, with the highest homicide rates in the states with an intermediate level of poverty. To actually have a model we can apply to data, we have to be more precise about how X and Y are related.

The E term, often called the **residual** or **error term**, suggests that things other than X cause Y to vary from observation to observation. Thus the equation indicates that Y is a function of X, plus an error term. In other words, our model says that poverty rates are not the only cause of homicide rates; other variables may be causes of homicide rates, and these are represented by the "E" in the equation. Just as Y and X can be different from person to person or state to state, E can also vary from person to person and state to state.

Let's pursue the example of homicide rates. Figure 1.1 graphs the homicide rate against the percentage of the population in each state in poverty. This graph is called a scatterplot because we are plotting the "scatter" of Y and X. We will discuss scatterplots in more detail in Chapter 5. The poverty data are for 2002, the homicide data for 2003. The homicide rate is the number of homicides per 100,000 population in the state. The poverty rate is the percentage of families whose income was

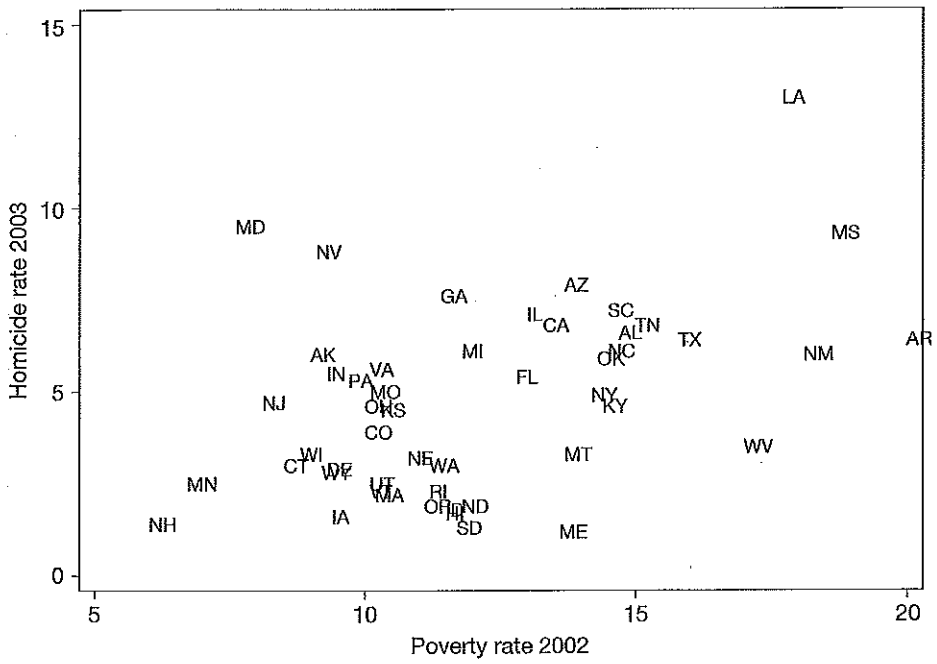

Figure 1.1 Relationship between homicide rate and percent of population in poverty
Data source: US Census Bureau (2002, 2003), analyzed with Stata.

less than the federal poverty line for families. The data are described in more detail in the applications at the end of the chapter. We usually want the independent and dependent variable to be for about the same point in time, and the custom is to have the independent variable be for a slightly earlier point in time than the dependent variable when they are not both for the same time.

Using words instead of letters to form the equation, the model we are proposing is:

$$\text{homicide rate} = f(\text{percent in poverty}) + \text{Error} \tag{1.2}$$

The model says that one reason that states vary in their homicide rates is that they have different levels of poverty. We don't know that the model is correct. Indeed as the quote from George Box (see Box 1.1) suggests, all models, including this one, are wrong. By wrong we mean that no model will predict the data perfectly. But as the quote from Samuel Karlin suggests, we can learn something from imperfect models. We do this by looking at the pattern in the scatterplot.

There does seem to be some tendency for the homicide rate to be higher in the states with the highest poverty levels. Note also that even though there is some pattern, it is far from perfect. The data points don't seem to fit any pattern perfectly even if there is something of a general pattern. States with the most poverty do seem to have the highest homicide rates. The deviations from the pattern are also interesting. States like Mississippi (MS) and Louisiana (LA) have higher levels of homicide than we might expect from their levels of poverty, while Maine (ME) and South Dakota (SD) have lower levels of homicide than we might expect from our model. For example, Arizona and Maine have about the same poverty level (13.5 and 13.4, respectively). But the homicide rate for Maine is only 1.2 while for Arizona the homicide rate is 7.9. So the poverty rate clearly doesn't predict the homicide rate perfectly. In the language of the model, the states that are different from the overall pattern will have large E values. Things in addition to level of poverty are having an influence on the homicide rate.

Box 1.1 Models

All models are wrong. Some models are useful. (George E. P. Box, 1979)

The purpose of models is not to fit the data but to sharpen the questions. (Samuel Karlin, 11th R. A. Fisher Memorial Lectures, Royal Society, 20 April 1983)

Both of these quotes from eminent statisticians remind us that models are tools to aid our understanding. As we develop models we can get lost in the modeling itself. We should always reflect back on the purpose of building and testing models: to help us understand the world. Sometimes a model that doesn't describe data very well tells us as much or more than a model that fits well.

To go beyond looking at a scatterplot, we have to be more precise about how X is related to Y. The shape of the relationship between X and Y is called the "functional form" of the model. In some software it is called the "linking function" because it is what links X to Y. The functional form or linking function is a very important part of the model, one that should be specified, at least tentatively, by theory. In the poverty/homicide rate example, we might suggest that the link is best represented by a straight line that indicates that as poverty rates go up, homicide rates go up. Then f(X) becomes the equation for a straight line:

$$f(X) = A + (B*X) \tag{1.3}$$

We read this equation as "F of X equals A plus B times X." When working with the equations, pay attention to where the parentheses are. In solving this equation you should multiply X times B, then add A. Here f(X) is the function that links X to Y. The result is a prediction of Y based on X, f(X) is the prediction of Y using X as the predictor. The equation f(X) is not the same as Y itself unless X can predict Y perfectly, which won't be the case with real data. So there has to be an E in the equation for Y itself. That is the equation for Y is now:

$$Y = A + (B*X) + E \tag{1.4}$$

While this is the simplest and most commonly proposed functional form linking dependent and independent variables, it is not the only possibility. There is nothing except more complicated algebra preventing us from saying that the link between X and Y is a curve rather than a straight line. As we suggested above, we might have the idea that states at a moderate level of poverty have higher homicide rates than those at the high and low extremes. This implies a curve that looks like an upside down letter U. We don't see this pattern in Figure 1.1, but it might occur for other variables. But it's best to start with the simplest models and the straight line is the simplest model we can have for the relationship between two variables.

Remember, E is still the error or residual term. It implies that we don't expect all variation in Y to be predicted by X when we assume that the relationship is a straight line. By including E we acknowledge that factors other than affluence may explain homicide rate. The inclusion of E in the model indicates that we don't expect poverty rates to predict homicide exactly – not all data points will fall exactly on the line.[2]

It is useful to think of even this very simple model as having four parts. First there is the *dependent variable*, Y, the thing whose variation from observation to observation (that is from person to person, state to state, country to country or year to year) we want to explain. In our example, this is the homicide rate. Second is the *independent variable*, X, the thing that we believe can, with its own variation, explain some of the variation in the dependent variable. In the example, this is the percent of the population below the poverty line. Third is the *functional form* (f) that links the independent variable to the dependent variable. For now we'll talk only about straight lines, but more complicated functional forms are often realistic. The key

point for now is that we are making a statement that says not only that we believe that the independent variable can predict the dependent variable but also that we will use a straight line to indicate the link between them. The fourth component of the model is the *error term* (E). E describes the difference between the actual value of Y and what we predicted using X. It takes account of the fact that X does not perfectly explain Y. The error term is not usually discussed by sociologists or other theorists, but it is a key issue for statistical theory. Indeed, as will become clear later, the meaning of any statistical procedure rests on the meaning of the error term, and statistical analyses are only as sound as our assumptions about the error term. So our four parts of the model are the dependent variable, the independent variable, the functional form that links them, and the error term.

The functional form, $Y = A + (B*X) + E$, suggests that a straight line describes the link between X and Y. But what values do we pick for A and B – that is, what line would we draw to represent the relationship between A and B? We could pick an A value and a B value by "eyeballing" the data. That is, we could take a ruler and draw a straight line through the graph and then use methods we learned in high school algebra and geometry to find the values of A and B for the equation from the line we drew. But that's not how we find the line in statistical analysis. Chapter 14 will describe in detail how we pick the line. For now, take our word for it that a good value to pick for A is 0.59 and for B a good value is 0.36. Figure 1.2 shows the graph where the line f(X), which for a straight line is A + (B*X), has been drawn in.

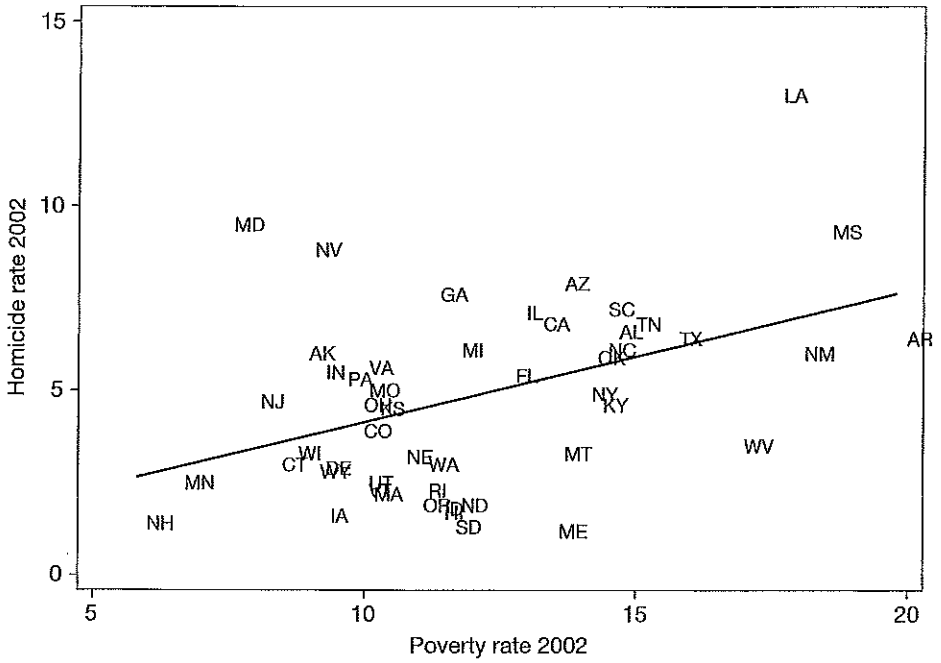

Figure 1.2 Relationship between homicide rate and percent of population in poverty with prediction line

Data source: US Census Bureau (2002, 2003), analyzed with Stata.

A is the value we predict for the homicide rate if there were no families living in poverty in the state – the situation when X equals 0. To see this, look at Equation 1.2. If X = 0 then B*X = 0. Then the prediction for Y is just A, which for this example is 0.59. Often the A term in a prediction doesn't tell us much on its own. Our real interest is in the B value. Remember the B value tells us how Y changes with X – how homicide rates change with levels of poverty. That's the whole reason for doing the analysis. In this case the A term doesn't mean much because there are not states at, or even very close to, a zero poverty rate. So in this example A just gives a reasonable line to describe the data. If the data included values of zero for X, then A would be more meaningful.

B says that for every 1-point difference in the poverty rate, the homicide rate increases by 0.36. (Remember that the poverty rate variable is a percent, so a 1-point increase is a 1% increase.) Let's look at what the model predicts for two states, Vermont and New York. For Vermont, the poverty rate is 9.9% so the model predicts a homicide rate of (0.59 + (0.36*9.9)) = (0.59 + 3.56) = 4.15. For New York, the poverty rate is 14% so the model predicts a homicide rate of (0.59 + (0.36*14)) = 5.63. Remember, these are predictions from the model. The predictions usually won't be exactly right. For example, the actual difference between the homicide rate for Vermont is 2.3 – 4.15 = −1.85. This means the model predicts too high a homicide rate for Vermont by about 1.9 points. For New York the difference is 4.9 – 5.63 = −0.73. Here the model predicts too high by nearly three quarters of a point. Of course, if we had picked other states, we would see that sometimes the model underpredicts.

Let's look at what the model is saying for a few states. Table 1.1 shows the values of Y, X, f(X) and E for a few states. For California, X = 13.1. We multiply this by B, which is 0.36 and get 0.36*13.1 = 4.72. Then we add A, which is 0.59, so (0.59) + 4.72 = 5.31. This is what we predict the homicide rate for California to be based on the poverty rate. The E for California is then the actual homicide rate minus the prediction. Then E = 6.8 – 5.31 = 1.49. We can see that for California the prediction was pretty close, but it underestimated the homicide rate slightly. For Nevada the prediction also underestimated the homicide rate. The actual rate was 8.8 and the predicted rate was 3.79, so the actual rate was about five points higher than what the model predicted. Note that positive values of E mean that the model *under*-estimates the value for that state, while negative values indicates that the model *over*estimates the homicide rate.

Table 1.1 Poverty rate, homicide rate, predicted value and error for selected US states

State	X_i *(% in poverty)*	Y_i *(Homicide rate)*	$F(X_i) = A + (B^*X_i)$	E_i
California	13.1	6.8	5.31	1.49
New York	14.0	4.9	5.63	−0.73
Nevada	8.9	8.8	3.79	5.01
Vermont	9.9	2.3	4.15	−1.85
Oklahoma	14.1	5.9	5.67	0.23

In Figure 1.2, we see the same thing by noticing that the line, which represents the model's prediction, is very close to the data point for California (CA), but pretty far away from the data point for Nevada (NV) and Vermont (VT). Underestimates mean that the state is above the line and the E value for that state will be positive. Overestimates mean that the state is below the line and the E value will be negative.

Explaining Variation

Variation explained is a central concept in many applications of statistics. Social scientists are concerned with why people, social institutions, communities, cultures, nations and other units of social analysis vary from one to another. We want to know why some people are rich and others are poor, why some nations have a high quality of life and others do not, why some social movements succeed while others fail. In the model just presented, we want to explain why states vary in their homicide rates. In most discussions, we'll actually talk about variance explained rather than variation explained. Both the terms "variance" and "variation" have precise definitions in statistics, as we'll see in Chapter 4. Until we get there and see the definitions, we'll use the term *variation* in its everyday sense – **variation** means that things differ from person to person or country to country.

It is important to remember that the term **explanation** in the context of statistics has a very precise meaning. In its technical usage, it refers to the ability to predict one variable based on another variable or set of variables. So in the example above when we say we can partially "explain" the homicide rate in terms of the amount of poverty in a state we mean the poverty rate can predict the homicide rate, though of course not perfectly. Explaining variation in this technical sense of *predicting variation* across observations can lead to explanation in a broader sense. This happens when the explanation of variation is linked to a practical or theoretical framework. Lacking such a framework, statistics cannot produce much understanding. We may develop models that have great explanatory power in the sense that they predict quite well but provide no theoretical or practical insight. Or we may learn a great deal about the social world from a model that has limited explanatory power but that reveals important patterns. Indeed, the fact that a model doesn't predict very well can lead to important insights, as the quotes in the box above indicate.

Quantitative methods are powerful tools, but they achieve that power only when combined with sound theoretical thinking. In contemporary social sciences, it is often the case that theorists pay little attention to issues of method and sometimes methodologists don't think about theory. Successful research requires making links between theory and method – between developing understanding of phenomena in a theoretical or practical way and explaining variation in a technical way. Indeed, this is the stage at which the craft of research is of paramount importance. Good

research makes connections between developing understanding of phenomena in a theoretical or practical way and explaining variation in a technical way.

Consider the relationship between homicide and poverty. Let's suppose we can predict the homicide rate with reasonable accuracy based on the percentage of each state's population in poverty. In the statistical sense we have "explained" a good bit of the variation in homicide. But if we do not have a theoretical model of the relationship between homicide and poverty, we have not really learned much.

We have to link the model to theory. For example, it may be that poor people are alienated and suffer from social disorganization and thus are more likely to engage in crime and violent activities. If we think that's the case, we might want to see if there is a relationship between poverty and other kinds of crime. This would support the idea that poverty leads to disorganization, which leads to crime though there still might be other things to consider. Or it may be that poor people are concentrated in cities and that homicide is more likely in cities simply from the volume of social interactions. Then we would want to look at the relationship between urbanization and both homicide and poverty. As we develop a theoretical explanation to work with our quantitative analysis, the theory often suggests other analyses with other variables. Looking at the data may answer some questions, but it always raises new ones.

Students learning the basic tools of statistical analysis sometimes mistake statistical explanation for theoretical explanation. As we will emphasize, effective use of quantitative methods requires a theoretical context for guidance. Quantitative methods can be a great aid in evaluating the ability of theories to inform us about the social world, but they cannot be used in the absence of theory. This is turn provides a strong challenge to theory. We are convinced that attempts to apply quantitative methods to our theories will greatly improve those theories. Quantitative methods force us to think hard about what we mean and what we assume.

The Use of Statistical Methods

Statistical methods are tools that aid in seeing the empirical world more clearly. Scientists observe phenomena, and from their observations they try to develop better understandings of the world. But many factors cloud and distort our observations, making it difficult to draw conclusions from them. Scientists, like all observers of the empirical world, see "through a glass darkly."

Statistical methods are designed to clarify that glass, to minimize the cloudiness, to help us sort "truth" from "error." Statistical methods cannot eliminate error, nor provide "truth," but they do provide an assessment of the magnitude of error that is there, and thus clarify our perceptions. In the equation presented earlier, E represents all the things in the world that may distort our understanding of the link between X and Y. It is all the things that prevent X from perfectly predicting Y. By understanding E we can improve our understanding of the relationship between X and Y.

Types of Error

There are several ways to think about the role of E – error – in scientific observation. We will discuss five ways to think about the use of statistics in the face of error: error in models, which keep models from predicting perfectly; sampling error; randomization error in experiments; measurement error; and comparisons of independent variables to random error. In this chapter, we introduce these ideas. But they will return again and again in subsequent chapters. We're sure that the more you see them the clearer they will become, so don't worry if the seem a bit difficult to grasp at first. Eventually, with enough understanding of statistics, you'll learn that the different kinds of errors are really different ways of thinking about the same thing.

Error in models

We've already looked at one example of a model – the simple model that uses the poverty level of a state to predict the homicide rate. This seems like a plausible idea, and the scatterplot in Figure 1.2 seems consistent with the argument. But we don't expect perfect prediction with our models. Not all the states have the values for homicide rate that the model predicts, in fact for most states the model is a bit off. This is not surprising – there must be more that causes homicide than just poverty, even if poverty turns out to be part of a good explanation. So our model of homicide rates has some error.

One way to think of the error is that it is everything else that causes the homicide rate. Some of those things are other variables on which we can obtain data, so we could expand the model to include those things, using methods we'll discuss in later chapters. But even if we include all the variables that theory suggests might be important, we still wouldn't expect the homicide rate to be perfectly predicted by the model. If our theories are good descriptions of the world, adding more variables will reduce the errors associated with the prediction for each state, but the error will never completely disappear for all states.

Consider a tragic example. If we had used homicide data for 1995, rather than 2003, the data for Oklahoma would have included the 169 deaths that resulted from the right wing terrorist bombing of the McMurra federal building in Oklahoma City. Such a tragedy cannot be predicted by a model of state homicide rates, though sociological analysis can lend considerable insight into terrorism. We never expect models to predict perfectly, but only to let us understand better how the world works. So our models will always have error.

Sampling error and randomization error are found in most applications of statistics. Before we discuss them, it is useful to introduce a concept central to much of statistical analysis – the arithmetic mean. You already are familiar with the mean but you know it by its everyday name – the average. As you know, the average score of a class on a quiz is calculated by adding up all the scores and dividing by how many

people took the quiz. Your personal average on all quizzes in a course is the sum of what you scored on all the quizzes divided by how many you took. A special case of the average is the proportion or percentage of people in a particular category. For example the proportion of people in a class who get an A is just the number who get an A divided by the number of people in the class. We will refer to the average in the next few examples, but you already know enough about it to follow the examples.

Sampling error

Many data sets are based on a sample of objects. We might do a survey of individuals or households. Or we might have a sample of things like states, cities, organizations, or nations. While data are available for a sample, our practical and theoretical concerns usually are with the population from which the sample was drawn. Survey interviews with 1,500 US citizens are of interest to the extent they lead to conclusions about the attitudes and values of all Americans. A researcher may have data on 100 school districts but would like to speak about all school districts. Researchers can analyze the data in hand, the sample, and draw conclusions about it. But in most circumstances we also want to generalize to the population from which the sample was drawn.

Why not collect data on everyone or every organization or every state? When there are a very large number of people or organizations in the population we are studying, the costs in time and money of getting data from everyone can be prohibitive. In fact, statisticians have shown that we often can get a better understanding of a population by being very careful about getting data from a sample rather than having the same resources spread very thin in trying to get data on everyone. Of course, sometimes we do have data on every unit, as in the case of our analysis of state homicide rates.

Suppose, as is often the case, that we are interested in a population average or percentage. For example, in one of our continuing examples, we will examine the percentage of people in a Ugandan national sample of adults who knew condom use can help prevent the transmission of AIDS. (We will discuss this example in more detail in the Applications sections at the end of every chapter.) In the survey 8,310 people answered the question. Of those 8,310, 6,420, or 77% said "yes" that condoms can reduce transmission. That's interesting, but what we really want to know is what percentage of people in the Ugandan adult population, not just the sample, would say "yes" to that question. The percentage in the sample is a guide, but we also expect it not to be exactly the same as the percentage we would get if we interviewed all adults.

We can think of the relationship between the sample percentage and the population percentage in terms of a simple model that looks like the models we have already examined.

$$\text{Sample percentage} = \text{Population percentage} + \text{sampling error} \qquad (1.5)$$

The sample percentage is what we calculate from the sample. We know it is 77 percent. We want to know the population percentage, but we shouldn't assume that it is exactly the same as the sample percentage. So we allow for the sample percentage to differ from the population percentage because of sampling error. For the right kinds of samples, statistical procedures allow us to learn a lot about the sampling error and thus about how our sample percentage may differ from the population percentage we would like to know.

The patterns in the sample may not accurately represent the population for several reasons. First, there is the problem of having a sample that was selected using a non-representative process, one that, intentionally or by accident, includes in the sample exceptional rather than typical cases. For example, we might use a convenience sample in which we interview the first 100 students we encounter in the student union. When we have a non-representative sample it is usually not possible to determine the relationship between what is seen in the sample and what is true in the population. We don't know how to generalize from the first 100 students we run into to the entire student body. With non-representative samples, statistics are of little help in going from the sample to the population. But graphical and descriptive statistical techniques can be used to understand the sample itself.

Generally, we would prefer to have a "representative" sample, but what does that mean? One way of thinking about **representative samples** is to have a sample constructed in such a way that every member of the population we're studying has the same chance of appearing in the sample. This is called an *equal probability of selection sample.*

Even with such a representative sample, conclusions drawn from the sample may not be perfectly accurate representations of the population because of chance processes. Sometimes an honest coin may come up tails for 10 flips in a row. In the same way, a random sample may display patterns that aren't typical of the population, purely by bad luck. If we drew names of students at random from the registrar's list, we might by luck get too many women, or too few chemistry majors, or some other non-representative mix. The advantage of a probability sample over a convenience sample is that in the probability sample, statistical procedures let us estimate the likely magnitude of the error produced by sampling. With convenience sampling the magnitude of the error cannot be known.

Ways of drawing probability samples are discussed at the end of this chapter as an Advanced Topic.

When the process by which the sample was drawn is understood, as is the case with simple random samples and other probability samples, then statistical tools make it possible to put probable upper and lower bounds on the errors generated in sampling. They don't eliminate sampling error, but they do indicate how large it may be, and can be used to place appropriate hedges on conclusions. This understanding

Box 1.2 What Are the Chances of Getting a Badly Non-Representative Sample?

What are the chances of drawing a sample composed entirely of men? Suppose we draw a sample of 100 students, from a student body (the population we want to sample) that is half men and half women from a very large university, so that 100 students is a tiny fraction of the study body. If we use a convenience sample by taking the first 100 people walking out of the student union, we can't calculate the chances of getting a sample of all men, but we might imagine things that would make that happen – perhaps we go to the student union right after a fraternity rush event ends. But if we draw the sample at random from the registrar's list of students, we can do the calculation of how likely it is to get all men in the sample. It's the same as the probability of getting 100 heads in a row when tossing a fair coin. The chances are about 0.0000000000000000000000000000001 or one in ten nonillion. Pretty small chances, not something we really need to worry about. In Chapter 7 we'll show you how to do these calculations.

is a fundamental insight in statistics, one that emerged in the late nineteenth and early twentieth centuries. Before that time, scientists tended to work with whatever data were available to them – a convenience sample. They had no clear sense of what data might be representative, and thus useful for generalizations, and what data might be misleading because it was not representative of a larger group. Now in most branches of science, careful attention is given to obtaining a representative sample of the things being studied.

Randomization error

Sociologists and most other social scientists, except psychologists, don't often conduct experiments. But we will discuss randomization error because it has played a central role in the development of statistical thinking and because in some fields like psychology experiments are very important sources of evidence.

When a simple experiment is conducted, the subjects are sorted at random into two groups, one labeled the **experimental group** and the other the **control group**.[3] In experiments with **random assignment**, the two groups are created by a chance process, such as the flip of a coin. "Heads" and you're in the control group, "tails" and you're in the experimental group.

During the experiment, something is done to the experimental group that isn't done to the control group. Then the two groups may differ from one another for two reasons. One is because of the factors manipulated by the experimenter – the thing done to the experimental group but not the control group. The other is the chance process by which subjects were assigned to groups. The power of the

experimental approach comes from its relatively unambiguous ability to attribute differences between groups to the experimental manipulation when it is not reasonable to believe that the differences between the groups were due to chance.

Suppose the two groups are created by the flip of a coin. The experimental group might watch a music video that shows stereotypical gender images.[4] The other group watches a music video that has no explicit gender content. After watching the films each group fills out a questionnaire that measures attitudes about gender relations. If the experimental group, on average, scores higher than the control group on items indicating a belief in adversarial gender relations, there are two possible explanations. One explanation is that the gender stereotyped video had an effect, relative to the "neutral" video. The other explanation is that the two groups had different attitudes at the start.

Since people were placed in the groups by the flip of a coin, statistics can assess the likelihood that the two groups differ in gender attitudes as a result of the coin flip that sorted them into groups. If the difference in attitude between the two groups is too large to plausibly attribute to chance then there seems no reasonable explanation except the argument that the film had an effect on the viewers.

Of course, the coin flip could have assigned all those with conservative gender attitudes to one group purely by chance. But statistical analysis that we will learn to do later says the chances of that happening are too small to be believed. So we have more faith in the explanation that the video content had an effect on attitudes.

On the other hand, if the differences between the two groups in gender attitudes were of the kind that flips of a coin could easily create, then the safest explanation may be that the video had no effect on attitudes. In the experiment, the probability that the differences were a result of sorting people into two groups was .021, allowing Kalof to conclude that exposure to the stereotyped video did have an effect on attitudes.

Again, a simple model can serve to explain what might have happened. We could calculate the average score on the gender relations scale for the experimental group and the average for the control group. Then if the video has no effect, the model would say that the two groups differ in their scores on the scale just by luck of the coin flip, which is random error. The model is:

$$\text{Average of experimental group} = \text{Average of control group} + \text{Randomization Error} \tag{1.6}$$

Remember, all the error we are dealing with can be positive or negative, so the control group could have an average higher or lower that the experimental group. Also, there's an equivalent way of thinking about this model. We can subtract the control group average from both sides of the equation. Then we have:

$$\text{Average of experimental group} - \text{Average of control group} = \text{Randomization Error} \tag{1.7}$$

Or to put it more clearly,

Difference between experimental and control group
= Randomization Error (1.8)

If this model explains our results well – that is if the difference between the two groups is the kind of thing a coin flip could generate – then we would conclude that the video had no effect. If the difference we find is not what we would expect as a result of random error, as was the case in Kalof's actual experiment that we are using as an example, then we would conclude the experimental treatment had an effect. So in a sense we are comparing the difference between the two groups to a random number and saying we have an effect if the difference is bigger than a random number would be.

Measurement error

Most scientific observation involves some mis-measurement of the variables being studied. Survey questions tap individual attitudes and values; official statistics provide information on economic, social and environmental processes in cities, states and countries. However skillful the designer of the survey, however honest the respondent, however careful the statistical office, errors inevitably will creep into the data. These errors in measurement may distort observed patterns.

In the example of homicide rates, we know the homicide rate for each state almost certainly has some error in it. Some homicides are never detected while some deaths that are not homicides might be misclassified as murders. Statistics can provide tools to minimize this error, and under the proper circumstances provide an estimate of how large such errors may be. In fact, many statisticians work on finding ways to reduce measurement errors in studies where the outcomes are important. The US federal government employs many highly trained and dedicated statisticians who are continually developing methods to better measure things like the population or the unemployment rate so that we can make decisions using the most accurate possible information.

Measurement error was central to the origin of statistical analysis (see Bennett, 1998: ch. 6). By the sixteenth century, astronomers and other scientists were noting that if they made the same observation over and over, they got slightly different results each time. In 1632 Galileo noted that:

- the errors were inevitable;
- they were mostly small with relatively few large errors;
- they were symmetric in the sense that overestimation was as common as underestimation; and
- the true value was in the area where most observations were clustered.

By the eighteenth century, a number of scientists and mathematicians had suggested that a bell-shaped curve described the measurement errors. They began to link the

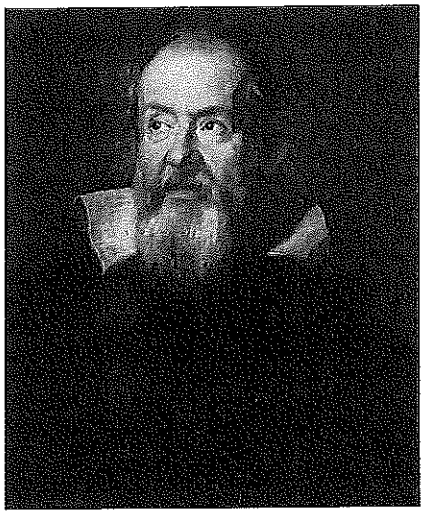

Figure 1.3 Galileo
Source: http://commons.wikimedia.org/wiki/Image:Galileo-sustermans2.jpg.

idea of errors to random processes. Thus the observation could be thought of as the true value plus (or minus) some measurement error that was generated by a process rather like tossing dice or drawing lottery tickets. With this idea the basis for modern statistics was developed – our observations of the world include random elements and that statistical procedures can help separate out the random elements and provide a better understanding as a result. In statistics, we call processes that generate such random error "stochastic" processes. The term stochastic is derived from the Greek term for "skillful at aiming or guessing." A simple model of this error would be:

$$\text{Measured value} = \text{Actual Value} + \text{Measurement Error} \tag{1.9}$$

The more we know about the measurement error, the better we are able to know what the actual value of the thing measured will be.

In the case of the homicide rate and poverty example, the idea of measurement error would suggest that if we had perfect measurement of homicide rates (that is if the E term for every state were zero) then we would perfectly predict homicide with poverty. This does not seem like the most reasonable way to interpret the E term in the model. We don't think that the model misses badly for Nevada simply because the measurement error for homicide rates is very large there, while it is smaller for California. There are better ways of thinking about the error in our example. But for many other problems, such as those where we are measuring attitudes or values, the measurement error interpretation of the E term is very helpful.

Perceptual error

The social world is complex, and available theories suggest many factors that are important. In addition, the behavior of individuals, groups, institutions, and nations are not rigidly determined, so it can be difficult to see patterns in data, or conversely, patterns that are not really there may seem to leap from the page or screen.

One value of statistical analysis is that it can provide powerful methods for arranging data, including graphs and summary statistics that make it easier to see patterns and to identify particular observations that deviate from the general pattern. Descriptive statistics and graphics help us minimize perceptual error. Over the last 20 years or so, some of the best minds in statistics have devoted much of their time to developing new graphs and summaries that reveal patterns in data.[5]

Comparison to random numbers

One way to think about all the sorts of error we have mentioned is to ask if the independent variable in a model acts any differently than a random number. We introduced this idea in discussing a model in which the difference between experimental and control groups is seen as just random error resulting from the process of assigning people to groups. Suppose we generate a variable by flipping a coin and recording head or tails (0 or 1) or by tossing a die and recording the number that comes up, or by having the computer generate a number picked at random. If an independent variable has as much effect on the dependent variable as a random variable generated by flipping a coin or tossing a die, then it's hard to argue that the independent variable has an important effect on the dependent variable. A study by the statisticians Freedman and Lane (Freedman and Lane, 1983) provides a nice example of this kind of logic.

Freedman and Lane were interested in the fact that at a prestigious graduate university, 28 percent of men applying in a given year were admitted, but only 24 percent of women. All other things being equal (an assumption to keep the example simple), is the difference in admissions rate real or accidental? Should the researcher believe that the admissions process is blind with regard to gender, or is there a reason to be concerned about discrimination?

The model might look like this:

$$Y = f(X) + E \tag{1.10}$$

In this case Y is whether an applicant was admitted. X is the gender of the applicant and E is the error term that suggests that gender may not perfectly predict admissions. Then the question becomes how to interpret E.

Feature 1.1 Tea Tasting and Random Numbers

Freedman and Lane draw on some ideas from R. A. Fisher. Fisher is one of the founders of modern statistics. His name will come up throughout the book. One famous example of this kind of logic – comparing a real variable to a random number – is the story of "The Lady Tasting Tea" which is apparently a true story (Salsburg, 2001, pp. 1–8). It seems that at a formal tea at Cambridge University in England in the 1920s, a woman claimed she could tell the difference between cups of tea depending on whether the milk was poured into the cup before the tea was poured or the tea was poured before the milk. R. A. Fisher was in the room and proposed an experiment. The woman left the room and several cups of tea were prepared, all in identical cups, some with the tea poured first, some with the milk poured first. The woman was then asked to taste each cup and state whether it was a "tea first" or "milk first" cup. If you guessed randomly by letting a coin flip determine your guess with heads being "milk first" and tails being "tea first" you'd get about half the cups correct. So for the woman to demonstrate that she could really tell the difference, she had to perform considerably better than a random process (a coin flip). Apparently she did, indicating that it's likely that she could tell the difference (or she was very lucky that day).

Freedman and Lane make the following argument. Suppose that, instead of gender, the researcher had assigned each person a score on a random number that has two values. That is, suppose each person seeking admission was assigned a score of "heads" or "tails" based on a coin toss. Then the researcher cross-tabulated that number with admissions, calculating an admissions rate for "heads" people and "tails" people.

Are the results for gender (with two theoretically meaningful categories, female and male) much different than those that come from the random variable with the non-meaningful categories of "heads" and "tails"? To put it more precisely, how often would a difference of 4% result from calculating admissions rate differences between "heads" people and "tails" people?

If the observed gender difference looks like a typical result obtained from the random "coin flip result" variable, it is hard to argue that gender was an important factor in the admissions process. It turns out that many standard statistical methods apply to such problems, including the chi-square method we will learn in Chapter 12. In this case chances are about one in four that a difference of 4% would occur if admissions rates were calculated based on the random variable. Thus we would conclude that nothing important is going on, and the observed difference may well be a fluke.

Comparison of the independent variable to a random number is one way to think about the error in our model of state homicide rates. The data are not from a sample. We have all of the data available. In other words, we have data for the entire population of 50 states. Nor were the values of the independent variable assigned

at random as they would be in an experiment – we haven't randomly assigned states to different poverty levels. While there may be measurement error, we are not comfortable saying that measurement error is the only reason (other than poverty) that states vary in the rate of homicide. So what does the E mean? We could ask the question of whether the poverty rate is behaving the same way that a random number would in its predictions of the homicide rate. Unless the poverty rate is a better predictor of the homicide rate than a random number, we wouldn't put give much credence to the theory that says poverty is a cause of homicide.

Another way to think about this situation is to imagine a **superpopulation.** If we have data on all the states in the US or all the nations in the world, we don't really have a sample in the conventional sense, we have a population. But we can imagine the US states having evolved a bit differently than they did, and thus would have different values on the variables we are studying than they actually have.

The same argument could be applied to the nations of the world. We may have the full set of nations, and in that sense we are studying a population – or at least the process that leaves some nations out of the data set cannot be considered random sampling. But we can conceive of a hypothetical population of nations with different mixes of values on the variables of interest. We call this population of all the nations or all the states that "might have been" a superpopulation.

Then we can think of random error as sampling error that gives us a particular set of values for the countries or states in our actual data sets (the ones from the world in which we actually live) just in the way that sampling error in a survey describes the way the mix of people in the sample may differ from the population. We treat the data we actually have as a random sample from the superpopulation.

The statistical procedures are intended to tell us whether what we see in our data are likely to also be true in the "superpopulation," just as statistics tell us about the likely ways a sample may differ from the population from which it was drawn. The difference is that in a survey, the population is real and with enough money and time we could do a survey of everyone, while the superpopulation is just an idea that helps us understand error.

To summarize, we always assume that our models contain error. The error may come from limitations of a model linking an independent to a dependent variable, sampling, from random assignment in an experiment, from measurement flaws or from comparing a real variable to a random number. Statistical methods can help us understand how large the error in our analysis might be. This allows us to make statements that take account of the error, and draw conclusions in the face of that error.

Assumptions

It is common in discussing statistics to mention that all statistical procedures make assumptions about the data, the population and the processes that generated the

data. Such discussions then note that statistical techniques are valid only insofar as the assumptions that underpin them are met by the data being analyzed. But rather than talking about assumptions being correct or incorrect, it makes more sense to talk about our models being more or less correct. Remember that the model contains a random element as well as the social variables we are studying. We have to ask if the random model is a pretty good description of the world.

That is, to use statistics we have to postulate a model, and our results depend on that model. If we are ignoring key factors, if the model badly misrepresents reality, then the conclusions we will draw from comparing the model with reality are likely to be wrong. The ability of statistical tools to separate error from truth and to estimate the magnitude of error depends on the random part of the model being roughly correct.

In some cases the random part of the models we use is not hard to justify because the process by which data were collected is well known and matches the conditions under which a technique works (the assumptions flow from these conditions). This is the case when we apply statistical tools to random samples and to data from experiments with random assignment to experimental and control groups. But in the case of measurement error and comparisons of real variables to random numbers, we have to be careful to think through what the results of a statistical analysis really mean.

So again we see there must be a constant and critical interplay between statistics and theory. When we consider various statistical procedures in later chapters, we will indicate the assumptions they presume. We view these assumptions as part of the model to be assessed critically. Sometimes a model may be very implausible because some of the underpinning assumptions seem unreasonable. In other cases, the assumptions might be a bit inaccurate, but we can still get a reasonable description of the world.

Feature 1.2 Diversity in Statistics: Profiles of African American[6] and Mexican American[7] and Women[8] Statisticians

There is a rich history of mathematics in Africa south of the Sahara, a fact that until recently was largely ignored by most historians of mathematics. Indeed even the black African roots of Egyptian mathematics are often denied or otherwise rendered invisible by Eurocentric views of both "history" and "mathematics." But in fact, the Greek civilization that we tend to revere in the history of mathematics was in fact a Mediterranean civilization more than a European one. Ideas were flowing back and forth from Greece, Italy, the Middle East and Africa. During the periods when Greece then Rome dominated the Mediterranean and on to the Middle Ages, European mathematics was "stuck" because there were religious and philosophical objections to the idea of zero and infinity. But during that period, Indian, Arab, and African mathematicians were making great strides because they had no religious objections to zero and infinity (Seife, 2000).

But in more recent times, science has been dominated by men of European origin, and statistics is no exception in this regard. Some of this is because of poverty and lack of opportunity, some of it because of outright discrimination. So as we move through the text and make reference to statisticians who developed the methods we use, it would be easy to get the impression that nearly all statisticians are men of European origin. But that is not true. Many important contributions to statistics have been made by women and people of color. We won't see all those people in the main text because statisticians often worked on topics more advanced than we can cover in the introductory book. To highlight their contributions to statistics, we provide brief profiles of three African American mathematical statisticians, a Mexican-American statistician, and two women statisticians.

David Harold Blackwell was the seventh African American to receive a PhD in Mathematics in the US (University of Illinois, 1941, and he was only 22 years old at the time). Dr Blackwell was also the first African American named to the prestigious National Academy of Sciences – the highest honor an American scientist can win short of the Nobel Prize. Dr Blackwell's most famous work was on game theory, and his book *Theory of Games and Statistical Decisions* (1954) is considered a classic in the field. He started his academic career at Howard University where he was promoted from instructor to Professor in three years. He then moved to the Statistics Department at the University of California at Berkeley where for many

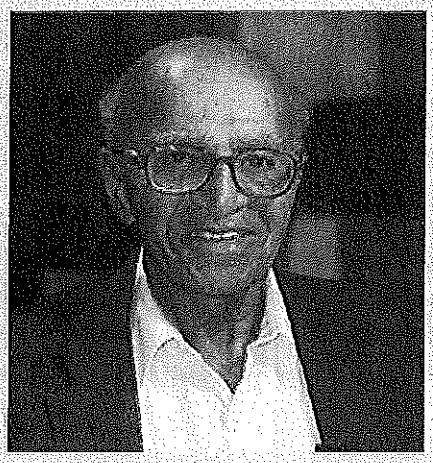

Figure 1.4 David Harold Blackwell
Photograph by Skip Coblyn; © 2008 National Visionary Leadership Project

years he was Professor and Chair of one of the most prestigious statistics department in the US.

Charles Bernard Bell, Jr, received his PhD from the University of Notre Dame in 1953. He has held appointments as Professor of Mathematics and Statistics at Case Western Reserve University, the University of Michigan in Ann Arbor, and Tulane University. Dr Bell has worked with mathematicians in Kenya and India, and has developed courses in mathematics for teachers in Nigeria. He is the author of numerous papers on nonparametric statistics and stochastic processes – that is on statistics that don't make many assumptions about the data and on random processes that are at the heart of statistics.

Albert Turner Bharucha-Reid studied at the University of Chicago (1950–53), after receiving his BS in Mathematics and Biology from the University of Iowa at the age of 19. He left Chicago before finishing his PhD (which he thought was a waste of time). He has held faculty appointments at the University of Oregon and Wayne State University in Detroit (as Full Professor and Dean of the School of Arts and Sciences). Bharucha-Reid has published more than 70 papers and 6 books, including *Probabilistic Analysis and Related Topics* (Academic Press, 1983), *Random Polynomials, Probability and Mathematical Statistics* (Academic Press, 1986), and *Probabilistic Methods in Applied Mathematics* (Academic Press, 1968).

Javier Rojo earned his PhD in Statistics from the University of California at Berkeley, and he is currently Professor of Mathematics at the University of Texas, El Paso. Dr Rojo and his four sisters grew up in Juarez, Mexico. His parents did not finish grade school. He was always good at mathematics, and after high school he earned a Bachelor's degree in Mathematics at the University of Texas at El Paso and then a Master's degree in Statistics at Stanford University before going on to Berkeley for his doctorate. Dr Rojo views statistics as a critical tool for better understanding the problems in society. Dr Rojo has examined the impact of the 20-year-old Clean Air Act on pollution in the national parks, and he has studied the mapping of genes in the human genome to determine whether a particular gene has an impact on one's chances for getting certain diseases.

Florence Nightingale David earned her PhD in Statistics at the University College, London, in 1938, after five years working under the guidance of Karl Pearson and Jerzy Neyman, two influential modern statisticians. During World War II she developed models of the effects of the German Blitz on London that were very important in allowing the government to save lives and provide essential services during the bombing. Dr David held faculty appointments at the University of California at Berkeley and the University of California at Riverside, where she was Professor and Chair of the Department of Biostatistics. Dr David authored more than 100 scientific papers and 9 books, including the classic book on the history of probability theory, *Games, Gods, and Gambling*.

Figure 1.5 Florence Nightingale David
Source: http://mathdl.maa.org/images/upload_library/1/Portraits

Figure 1.6 Gertrude Mary Cox
Source: North Carolina State University Libraries.

Gertrude Mary Cox earned Bachelor's and Master's degrees at Iowa State University with the support of George Snedecor, one of the pioneers in using statistics to analyze experiments. She then went to the University of California at Berkeley to earn her PhD in statistics. Snedecor encouraged her to return to Iowa State, and together they developed tools for analysis of experiments that are still in use. Her classic book with Snedecor, *Experimental Designs* (1950), is cited in the research literature hundreds of times each year. When Snedecor was asked to recommend someone to start the statistics department at North Carolina State University, he listed ten men and then added a line: "These are the ten best men I can think of. But, if you want the best person, I would recommend Gertrude Cox." Cox went on to become a Professor of Statistics at North Carolina State University at Raleigh and the founder of one of the strongest statistics departments in the US. The "Research Triangle" area of North Carolina, where North Carolina State, the University of North Carolina and Duke University are located is a major center for statistical work, and much of this can be traced to the leadership of Cox. She was the first woman elected into the International Statistical Institute in 1949 and was named to the prestigious National Academy of Sciences three years before her death in 1978.

What Have We Learned?

Quantitative analysis proceeds by developing models to explain the variation in one variable, the dependent variable. We do this by finding one or more independent variables whose variation we believe causes variation in the dependent variable. Thus quantitative analysis builds models to explain variation.

Statistics provides the tools for building and assessing models. Of course, we don't expect that any model will explain the world perfectly, so all quantitative models include error. This error is as important as the rest of the model. There are a number of things that can generate error in our models.

One is sampling error, the fact that data that are a sample from a population may not perfectly represent the population. Another is randomization error in experiments where people are sorted into experimental and control groups through a random process like the flip of a coin. For both sampling and randomization error, if we understand the process by which the sample was drawn or by which people were sorted into experimental and control groups, statistics can tell us how large the error is likely to be and give us a sense of when we are seeing valid results and when what we see is probably just error.

In some cases we have data on whole populations, and there is no experimental assignment. Then the interpretation of random error in our models is more subtle. Sometimes it's helpful to think of the error we encounter as measurement error. At other times the error takes the form of comparing a variable of interest to a random number. But whatever we are studying, in order to use statistical tools, we have to understand what may have caused random error to influence our data. Once we have thought through the origins of error in our data, we can use statistical tools of the kinds we'll learn in the following chapters to understand our data despite those random errors.

Advanced Topic 1.1 A Third and More Technical Definition of Statistics

There is a third definition of statistics that is more technical than the other two. As noted in the first definition, one of the primary uses of statistical methods is to draw conclusions ("make inferences" in the language of statistics) about a population based on data from a sample of that population. For example, we often have a survey of the population of the US and want to use the survey to estimate (guess) at what people in the whole population think. It is common to refer to the numbers we calculate using the sample data as the sample statistics.

We can think through this process by considering the Demographic and Health Surveys (DHS) (www.measuredhs.com). The DHS coordinates nationally-representative household surveys with over 75 countries on health, nutrition, and HIV-related topics. In 2000, a sample of Ugandan citizens were interviewed about their knowledge of AIDS, including whether they think condoms can help prevent the transmission of AIDS.

In the survey 6,420 people said "Yes" (correctly answered the question), 819 people said "No," and 1,071 people said "Don't Know." Since 8,310 people answered the question (6,420 + 819 + 1,071), the percentage saying yes in the sample is 6,420 divided by 8,310, which equals about 77 percent. We know this number is true for the sample. In the technical language of statistics, this percentage is called a "sample statistic." A *sample statistic* is just a number that is calculated to describe the sample.

The parallel numbers in the population (the percent in the whole population who would correctly answer the question) are then called *population parameters*. So if we had data on the whole Ugandan population and calculated the percentage saying "Yes" we would have the population parameter. We make this distinction because we know, as a result of simple arithmetic, the sample statistics. But we don't have data on the whole population so we don't know the population parameters. An important part of statistics is learning how to use data in the sample to make good estimates (guesses) of the population parameters. Statistical methods can tell us how to use the information that 77 percent of the sample answered correctly the question about how AIDS is transmitted.

Advanced Topic 1.2 Ways of Drawing Probability Samples

We need a "probability sample" to use statistical techniques to understand the size of sampling error and be able to use sample information to construct good guesses of what is true in the population. A probability sample is one in which we know the probability (the chances) that each and every member of the population ended up in the sample. The simplest kind of probability sample is the "simple random sample" in which every member of the population has the same chance of being in the sample. The statistical formulas for handling sampling error in simple random samples are simpler than those for other kinds of probability samples. We will only present the formulas that go with simple random samples in this book.

But sometimes it isn't practical to do a simple random sample. Suppose we want to compare two groups of very unequal size – say a majority group that is 90 percent of the population and minority group that is 10 percent of the population. If we draw a simple random sample of 500 people we'll have about 450 members of the majority group and only about 50 members of the minority group. We have a representative sample of the population,

but our comparisons between groups will be limited by the small number of minority group members. We might be better off by intentionally *oversampling* the minority group. We might design the sample so that we get a simple random sample of 250 people from the majority group and a simple random sample of 250 people from within the minority group. We can do this by drawing a simple random sample of each group. This gives us a much better ability to draw conclusions about the differences between groups than would a simple random sample of the whole population. And as long as we know how the groups are split in the population (90–10 in this example) then we still have a probability sample. This approach is called a **stratified sample**. We can still handle the sampling error involved but the formulas are a bit more complicated than for a simple random sample.

Another practical constraint leads to what is termed a **cluster sample**. It is often a good idea to collect data with face-to-face interviews. Sometimes we have to visit the offices of an organization to get data from their files. For example, many studies of the criminal justice system draw samples from police or court documents. But if we draw a simple random sample of people or organizations from a large geographic area such as the whole country, then interviews and site visits are scattered about, literally at random, all over the country. The time and travel costs involved gets very high.

As an alternative, some study designs draw the sample in stages. For a survey we might first pick counties within the US. We could set the sampling up so that each county has a chance of being picked that is proportional to its population. Thus counties with very large populations are almost certain to be picked in the sample while counties with very small populations have very little chance of being in the sample. Then we might pick blocks within the county, again with the probability that a block is picked proportional to the population of that block.

When we have finished this process, we will have a sample in which the probability that any member of the population is selected is known but where interviews will be clustered in a reasonable number of areas. Again, there are statistical procedures that allow us to generalize from a sample drawn in this way to the population, but the formulas for those procedures are a bit more complicated than those we use with a simple random sample. We will discuss the ways of drawing samples again in Feature 7.4.

Applications

Some students learn best by focusing on the theory first, others do best by beginning with examples. Even for the students who prefer the theory, examples provide a check on their understanding. In the text, we will use several extended examples that will come up in each chapter. This will allow you to build on a base of prior understanding rather than encountering each example without a starting point. Each of these examples is based on a research question that appears in the literature and a data set that can be used to try to answer that question. Here we introduce the examples.

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

We have begun to explore this question in this chapter. As you can see from the graphs we've used in the chapter, homicide rates are higher in states with high levels of poverty than in other states. By looking more carefully at the homicide rates, we can see that high homicide states are clustered in the southern part of the United States. This might be because that's also where the high poverty states are located. But the literature on homicide offers other explanations.

One theoretical explanation for the high homicide rates in the south is that there is a culture of violence in that region of the country that promotes homicide, driven in part by its history of slavery and lynching and a widespread use of guns (Baron and Straus, 1988). Another theory argues that the high rate of homicide in southern states is the result of the high rate of poverty in the region. Thus, poverty and economic inequality and their links with social disorganization cause homicide.

There is, however, a third line of reasoning that might explain why there are more homicides in the south than in other regions of the country: environmental conditions. For example, research has documented the connection between aggressive behavior and increases in temperature (Anderson and Anderson, 1984; Harries and Stadler, 1983). The high homicide rates in the south might be due to the hot climate of the region.

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

Animals are of substantial importance in human society. It has been theorized that research on how humans regard other animals, or their degree of concern for other animals, provides important information about how we organize our social worlds and how we see our connection to other living things (Arluke and Sanders, 1996). Thus, understanding variation in animal concern may provide us with insights into human character.

We could first look at differences between women and men in their concern for animals. Examining variation in animal concern by gender is based on the theoretical argument that women are more likely than men to be caring and to make moral decisions based on an ethic of care (Gilligan, 1982). Thus we would expect that women would be more likely than men to be concerned about animals.

Another argument suggests that an individual's concern for animals is rooted in a connection the individual makes with the oppression and exploitation of other living organisms. Thus, the experience of oppression produces empathy for other oppressed individuals, human and nonhuman. According to this line of thinking, women would have higher levels of concern for animals than men because of their experiences with oppression and exploitation. Minority groups that have been subject to discrimination would also have higher levels of animal concern.

Example 3: Why are some countries more likely to participate in environmental treaties than other countries?

It's widely accepted that the world faces severe environmental problems. Burning fossil fuels, like gasoline and coal, has led to a build-up in the atmosphere of

"greenhouse" gases that are changing the earth's climate. Species of plants and animals are going extinct at one of the fastest rates in the history of life on earth. Humans use about half the freshwater that flows through the earth's rivers and lakes.

In seeking solutions to environmental problems, treaties have been an important way for countries to make international promises to address global environmental problems. Nations differ quite a bit in their responses to environmental treaties. Some nations are quick to ratify most environmental treaties, other nations ignore them completely, and some nations are selective in their participation in treaties. Numerous factors, besides the merits of particular treaties, have been proposed to explain differences in environmental treaty participation among nations. (Dietz and Kalof, 1992; Frank, 1999; Roberts, Parks, and Vasquez, 2004)

One potential explanation is that international relationships play a role in treaty participation, with nations being encouraged to share common global values, including environmental protection, by other nations, especially in the context of membership in international organizations. Pressure by citizens, environmental movements, and other organizations to make environmental commitments may also be important. It also has been argued that nations with strong democratic institutions will be more likely to ratify treaties, especially in nations in which political accountability to citizens is high (e.g., politicians need to worry about being reelected and therefore want to be responsive to the public's interests). Additionally, countries that are environmentally vulnerable may have more incentives to participate in environmental treaties.

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

It has been said that since the bubonic plague in the fourteenth century, which killed about one-third of the population in Europe alone, no epidemic has had as strong an impact on population growth in the world as HIV/AIDS. According to the World Health Organization (WHO) and the Joint United Nations Programme on HIV/AIDS (UNAIDS), over 34 million people are living with HIV today (2000 statistic). And an estimated 18.8 million people in the world have died from AIDS since the beginning of the epidemic. Knowledge about HIV/AIDS prevention has been and will continue to be the key to reducing the spread of the disease (Population Reference Bureau, 2001; UNAIDS, 2000; World Health Organization and UNAIDS, 2006).

In 2000, 72 percent of the people in the world with AIDS were living in Africa. Uganda, located in East Africa, was one of the first countries to experience the HIV/AIDS epidemic. While rates of HIV/AIDS have been rising in most African countries in the past few decades, Uganda is one country where rates have been significantly declining. This success has been attributed to the nationwide effort – on the part of the government, non-governmental organizations, religious leaders, and community groups – to increase HIV/AIDS knowledge among its citizens. Efforts at improving education have been concentrated in schools via sex education

programs and in radio programs. Since the 1990s, there has been a large push to increase citizens' use of condoms, to educate individuals who have other sexually transmitted diseases in particular, and to increase the availability of HIV/AIDS tests.

In Uganda and worldwide, the AIDS pandemic has become feminized since the 1990s, meaning the HIV/AIDS virus is increasingly found among women (in sub-Saharan Africa today, around 60 percent of those testing positive for HIV are women). Furthermore, especially in developing countries, rates of HIV/AIDS infection have been increasing among married women. One reason is that many married women are at risk because they are not using any contraception during sex. Since knowledge is a key to reducing the transmission of HIV/AIDS, it may be that women and married people (married women in particular) have less knowledge about how HIV/AIDS is transmitted. Furthermore, it is reasonable to expect that people with higher levels of education are more likely to be knowledgeable about how HIV/AIDS is transmitted. Finally, the extent of HIV/AIDS knowledge may vary among those living in urban and rural areas. Urban dwellers have greater access to sources of education, including schools, radio programs, and health care facilities. While this example will draw on Ugandan data, it is applicable to other countries, including the United States and Europe, where rates of HIV/AIDS infection have been growing.

References

Anderson, C. A. and Anderson, D. C. 1984. Ambient temperature and violent crime: Test of the linear and curvilinear hypotheses. *Journal of Personality and Social Psychology* 46, 91–7.

Arluke, A. and Sanders, C. R. 1996. *Regarding Animals.* Philadelphia: Temple University Press.

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

Bennett, D. J. 1998. *Randomness.* Cambridge, MA: Harvard University Press.

Blackwell, D. and Girshick, M. A. 1954. *Theory of Games and Statistical Decisions.* New York: John Wiley & Sons.

Box, G. E. P. 1979. Robustness in the strategy of scientific model building. In R. L. Launer and G. N. Wilkinson (eds), *Robustness in Statistics,* New York: Academic Press.

Brown, L. 1993. *The New Shorter Oxford English Dictionary on Historical Principles.* Oxford: Clarendon Press.

Cleveland, W. S. 1993. *Visualizing Data.* Summit, NJ: Hobart Press.

Cleveland, W. S. 1994. *The Elements of Graphing Data.* Summit, NJ: Hobart Press.

Cox, G. M. 1950. *Experimental Designs.* New York: John Wiley and Sons.

David, F. N. 1962. *Games, Gods and Gambling: The Origins and History of Probability and Statistical Ideas from the Earliest Times to the Newtonian Era.* New York: Hafner Publishing Company.

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

Frank, D. J. 1999. The social bases of environmental treaty ratification, 1900–1990. *Sociological Inquiry* 69 (Fall), 523–50.

Freedman, D. A. and Lane, D. 1983. Significance testing in a nonstochastic setting. In P. J. Bickel, K. A. Doksum and J. L. J. Hodges (eds), *A Festschrift for Erich L. Lehmann,* pp. 185–208, Belmont, CA: Wadsworth.

Gilligan, C. 1982. *In a Different Voice: Psychological Theory and Women's Development.* Cambridge: Harvard University Press.

Harries, K. D. and Stadler, S. J. 1983. Determinism revisited: Assault and heat stress in Dallas, 1980. *Environment & Behavior* 15, 235–56.

Kalof, L. 1999. The effects of gender and music video imagery on sexual attitudes. *Journal of Social Psychology* 139, 378–85.

Population Reference Bureau. 2001. *2000 World Population Data Sheet* (www.prb.org).

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.

Salsburg, D. 2001. *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century.* New York: W. H. Freeman and Company.

Seife, C. 2000. *Zero: The Biography of a Dangerous Idea.* New York: Penguin Books.

Tufte, E. R. 1982. *The Visual Display of Quantitative Information.* Chesire, CT: Graphics Press.

Tufte, E. R. 1990. *Envisioning Information.* Chesire, CT: Graphics Press.

Tufte, E. R. 1997. *Visual Explanations.* Chesire, CT: Graphics Press.

UNAIDS. 2000. *Report on Global HIV/AIDS Epidemic, 2000* (www.unaids.org).

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).

Williams, S. W. no date. Mathematicians of the African Diaspora (www.math.buffalo.edu/mad).

World Health Organization and UNAIDS. 2006, December. AIDS Epidemic Update (www.who.int/hiv/mediacentre/2006_EpiUpdate_en.pdf).