

P A R T O N E

Introduction and Fundamental Ideas



CHAPTER 1



Statistics as Science

1.1 What You Will Learn in This Chapter

The study of statistics is unique. It provides the methodology for decision making; it provides the methodology for the evaluation and advancement of science. Yet it is a science itself. We will see that the primitive notion of **randomness**, or unpredictability, is pervasive in the observable world. Examples of randomness include weather variations, changes in stock market prices and volumes, interest rates, the sex of offspring, roulette wheels, atomic decay, minor oscillations in the earth's orbit, and tidal flows. All these examples involve variations that to some extent are unpredictable; the unpredictable component is called *random*. The interaction between all the sciences and the study of statistics is fundamental because randomness is inherent in all science; this arises from the inherent presence of randomness, inevitable errors in observation and measurement, and lack of complete control of any experiment. Because randomness is universal, all practical decisions must in some fashion discover solutions to decision making under uncertainty. Will it rain, or not? Will interest rates rise, or not? Will I get prostate cancer? In each case, what do I do and how do I decide? In broad terms, decision making under uncertainty is the subject of this entire book.

1.2 Introduction

We all know how to make decisions. Or do we? What criteria do we use, what facts do we need, and how do we assess those facts? When your doctor tells you that the probability of your operation being successful is at least 90%, what does he mean and how do you interpret that information? In any case, where did the number 90% come from, and can we trust it? What is the probability of getting cancer by sunbathing? How do I evaluate the answer when I get it? How does it compare to the probability of dying in an airplane crash or in an automobile accident? If the quoted probability is 1 in 100,000, is that big or small? How do I judge? Besides, what is this word *probability*, and what does it mean anyway?

When an atmospheric scientist tells us that there is global warming, that is an indisputable fact. Or is it? "Tax the millionaires," cries a political candidate, "and we

can balance the budget and increase welfare payments.” How plausible is that statement? Are there really enough millionaires to make this claim a feasible policy option? Aerosols and air conditioners are depleting the ozone layer, claims an environmental group; but how do they know and how accurate is that information? Democrats protest that under Reagan the rich got richer and the poor got poorer, whereas others with equal conviction claim the opposite. Why is this problem so difficult? Surely, this is a simple fact to determine. Or is it?

1.3 Statistics: A Framework for Decision Making

All these examples involve the making of decisions and the evaluating of claims by others. The decisions and evaluations all involve collecting and assessing data and determining their relevance to our problem. Data are the “facts” that we deal with in statistics. Data can be “official government statistics” that are produced by some governmental department or agency, such as measures of interest rates, price levels, or levels of productive activity. Data might be the information contained in surveys or opinion polls; or data might be observations on some experiment.

We might conclude from the tone of the previous two paragraphs that all data are not created equal. Some data may be more informative than others, some data may be less reliable than others; but how do we judge? What clues should we look for to discover that all may not be as stated, or that the data do not support the proponents as they claim? How much uncertainty can we tolerate; how much must we live with?

There is currently a raging debate about global warming; however, you may remember that only a little over a decade ago, we were being bombarded by fears of global cooling. Whatever is the truth of the hypothesized relationship between human activity and global warming, the debate, if it is to be decided on rational grounds, is an argument about atmospheric statistics and their interpretation. Without the concepts provided by probability and the theory of statistics, it would be difficult even to conceptualize the problem. Consider just one small example of this.

There is a positive association between the average temperature of the Pacific Ocean and the level of carbon dioxide in that region. Does this prove the claims of global warming? No; for the issues are more subtle than can be decided on the basis of casual empiricism. You will soon learn that empirically discovered patterns of association—*correlation* is the technical term—are not to be confused with causality. **Causality**, you will recall, is the concept that there is a logical link between two events that indicates that variation in one leads to changes in the values taken by the other. Although there are a number of hidden subtleties in this simple definition, the primitive notion is clear enough.

In our current example, we can immediately ask which of two plausible causal mechanisms is applicable. One claim is that high levels of carbon dioxide lead to warming; another is that the Pacific Ocean is an enormous carbon dioxide “sink”—the cooler the ocean the more carbon dioxide it can hold and the warmer the ocean the less it can hold. The former hypothesis claims that the direction of causality is from carbon dioxide levels to warming, and the other hypothesis is that the direction of causality is the opposite. Both involve a positive association between carbon dioxide levels and warming, but the policy implications are vastly different. To resolve even this question requires some fairly sophisticated statistical techniques.

But you may ask why should you have to worry about the matter, because that is why people hire statisticians anyway and you, for one, have absolutely no desire ever to become a statistician. Recognize however, that as a voter you are and will be asked to pass judgment on the matter by voting. In any event, the outcome is going to affect your income and lifestyle in a very significant manner. If the “global warmists” have their political way, you are going to be asked to pay a very heavy price in terms of lost income, lost freedom of choice, and a lower standard of living; “saving the environment” does not come cheap. If the global warmists are correct and we ignore the problem, we will also pay a heavy price in terms of lost income and a lower standard of living. Here is a difficult choice that will not go away with wishful thinking. Given the enormous potential costs, might not an investment in knowledge about the decision be helpful? But what knowledge and how should it be used? We will see that the field of statistics helps us to discuss this issue as well.

Continuing our example, let me illustrate the knowledge difficulty that someone wishing to vote responsibly has. In the following quote from the *Wall Street Journal* (July 11, 1996), Dr. Ellsaesser comments on the evidence cited by the Intergovernmental Panel on Climate Control (IPCC). Dr. Ellsaesser retired from the Air Force after 20 years as an air weather officer and from the Lawrence Livermore National Laboratory after 23 years in atmospheric and climate research. As you will observe, you will require a fair amount of understanding of statistical matters to be able to discern what the argument is about and who is likely to be more correct. Consider the quote:

By concentrating on IPCC rules and procedures, IPCC writers and supporters have managed to avoid the more important scientific debate as to whether the balance of evidence suggests that there is a discernible human influence on global climate.

Santer et al. attempted to identify this human influence by correlating year-by-year global patterns . . . of mean annual departures from the control run averaged over the last 20 years of each equilibrium experiment. An upward trend in the correlations of the annual means from about -0.2 ca. 1950 to about $+0.4$ ca. 1970 was uncritically accepted as confirming that the observed temperatures were evolving toward the temperature pattern predicted by the model. . . . Completely ignored were the facts that the annual correlations decreased from about 0.28 in 1910 to about -0.2 ca. 1950, and that while they remained relatively steady from 1970 to 1985, they decreased sharply after 1985, ending up in 1993 near 0.12, well below their starting point in 1910.

In first reading this quote, you can be forgiven for reacting that it is all unintelligible and you will never be able to fathom what is really being said. However, you should not be so pessimistic; after finishing this text you will be able to assess the statistical merits of this discussion and have a far better appreciation of what the debate is all about. You will also be able to recognize the inherent difficulties that are involved in such research, especially the subtle issue of what is, or is not, causing global warming, if indeed there is such warming.

AIDS has captured the public’s attention as no other disease has done in nearly a century. The theory and practice of statistics permeates this debate, at least at the technical level and should at the policy level as well. Consider an example chosen for its seeming irrelevance to a statistical approach. Whether people who have tested HIV positive should be allowed to pursue jobs that bring them into intimate contact with the public is a thorny issue because it brings two “rights” into conflict. On the one

hand, individuals seek the right to pursue their own goals, and the government has passed much legislation to facilitate a nondiscriminatory job market. On the other, people also have the right not to be needlessly exposed to life-threatening risks in the normal course of living, and the government has passed much legislation to support this right as well. The issue cannot be resolved in the absence of a knowledge of the risks and the relative costs that are involved, and this is a problem in statistics. If the risk of getting AIDS from one's dentist were zero, few would argue about restricting dentists with AIDS from practicing. But if it were absolutely certain that attending an AIDS-infected dentist led to getting AIDS, then few would object to preventing such dentists from practicing as usual. The reality is that the probability of getting AIDS from one's dentist is between these two extremes. The question now becomes how to trade off the cost of inhibiting the freedom of individuals to practice dentistry with the cost of infecting the population with AIDS. The theory of statistics provides not only the procedures but the very language for framing an answer to this question.

A related example that was posed in terms of medical ethics was debated in the *New York Times* (September 18, 1997). The ostensible issue was ethical, but the proper formulation of the problem requires some knowledge of statistical theory and practice as well as a good understanding of the rationale for experimental design. The National Institute for Health and the Centers for Disease Control and Prevention sponsored research on the transmission of AIDS by mothers to their children. The objective was to find an inexpensive method to prevent the transmission of HIV to babies that would be economically feasible in the developing world. The controversy arose because a very important part of the research strategy in trying to learn from experiments is the need to have a control group that does not receive the drug. This is important because in real experiments there are many other factors at play that can mislead researchers in measuring drug effectiveness. Further, it is imperative for a successful research design strategy that the recipients and the nonrecipients be chosen "at random"—that is, chosen so as to attempt to avoid unforeseen biases that might arise in the analysis.

The best strategy requires that not only the potential recipients do not know who gets the drug and who does not, but the monitors of the experiment do not know either. However, *if* the drug is successful and does not have severe deleterious side effects, the nonrecipients will have a higher incidence of HIV than the recipients. Please note the all important "ifs" in this premise. In analyzing this situation and passing judgment on its ethical status, one needs to understand the roles of controls and "randomization" in research; with them one can determine the effectiveness of a drug much more quickly and efficiently. Consequently, if the drug is in fact successful, the sooner we discover that fact, the sooner future babies will be saved. One needs to know how to balance the lives of future infants against those in the experiment. Of course, we have begged the question, because in these situations, we do not know whether the drug is effective, and we certainly will not know whether the drug has serious deleterious side effects. Imagine, for example, the reaction to this experiment if there were a substantial and horrifying side effect; one's position on the ethics of excluding some from receiving the drug might well be very different.

A tragic example of the need for the services of a competent statistician examining all the available data was provided by the explosion of the space shuttle *Challenger* in 1986. As we shall see later in the text, the available data clearly indicated a low-temperature problem with the O rings. The tragedy was that these data were not analyzed properly, and as a result the space shuttle exploded.

Enough of fire, famine, and pestilence, let us proceed to less-depressing subjects for statistical analysis. Your passion might be literature or art, so what can statistics do for you here? You may be thankful to know not much, at least directly, but the use of statistical procedures is creeping into the evaluation of art and literature. This is especially so in the detection of fraud and in the attempts to ascribe a text or work of art to a specific author or artist. The question is always, “How ‘probable’ is it that this book, or painting, was indeed the work of a particular person?” The statistical analysis of writer’s styles is now an elaborate specialized field. Later, we will discuss the role of statistics in analyzing the sales and popularity of rock music; so even here the “most dismal of sciences” has a niche.

How about “lifestyle” decisions? Are they the proper subjects of statistical thinking? Where does statistics come in, for example, in your choice of profession? Certainly, one will want to weigh to some extent the anticipated future earnings, even if one claims not to be “money hungry.” Other aspects involve the job’s implied life expectancy. A miner must face a higher probability of death from accident and work-induced diseases, but what is the income trade-off and how do you calculate it? How risky is it really to be a policeman? Is it worth taking 2 years off to get an M.B.A.? If I want to become an actress, how do I decide when to quit trying? Or finally, can I live on my inheritance, or must I learn a profession?

Other decisions are more personal. As a woman, if you delay marriage and having children for your career development, how do the probabilities for birth defects change? Suppose that you are contemplating a second career after successfully completing a first; how much educational investment is it worth? You are contemplating an operation for the removal of a tumor; the decision hinges on the probabilities of whether the tumor is benign and whether you will be seriously debilitated by the operation itself. Of equal importance in this case is how reliable the probabilities are that you have been quoted and how you take such knowledge into account.

A recent article, “The Fat’s in the Fire, Again,” in the *New York Times* (January 11, 1998) illustrates the need to understand the process of statistical analysis. Researchers analyzed the behavior of 324,135 adult males over 12 years and concluded that the excess risk of dying associated with obesity was modest and declined with age! Given all the previous evidence this was a remarkable statement, and as you might expect it generated a storm of controversy. The article ended with the following statements:

“The only disagreement is among people who don’t know the facts and don’t have the scientific evidence,” Dr. Manson said in an interview.

Not surprisingly, others disagree about the disagreement.

In the same article, other evidence was quoted that indicated that obesity was very highly correlated with diabetes, high blood pressure, and high levels of blood cholesterol—all potential killers. As you read this text, you will gain some insight into this controversy and recognize the impact of the vagueness of the research question, that correlation and causality are very different concepts, and that there are many questions about the experimental procedures that need to be asked before coming to the conclusion that obesity is a benign state.

You may worry about the safety of cars or other products; the levels of pollution, what is tolerable and how to decide; how much cholesterol you can consume; or the fact that your parents both had early heart attacks, the probability of your having a heart attack before 40, and how likely it is that your children will have early heart at-

tacks. You are offered an insurance annuity; is it a good deal? You just bought a new appliance; should you also purchase the maintenance contract? Is it worth getting a cheap airfare before you are sure that you will use it?

Each of these examples requires the collection of facts. We are inundated by data and facts. Everyone who is pushing a cause, or fighting those who are, claims to have the “facts.” But as we have already indicated, not all data are created equal. One of the tasks of statistics is to provide procedures for evaluating the value of any data. We have to ask whether the data that we are provided to buttress an argument really do buttress the argument. As we shall see, the way in which data are collected and processed for our purposes can change the meaning that we attach to the data and how we interpret the results of research.

These are just a few scattered examples of how a knowledge of statistical concepts will help you to pose the right questions and indicate to you the type of data that you might need to answer the questions. If nothing else, a knowledge of statistical reasoning will provide you with a framework within which you can analyze the problem you face.

1.4 Statistics and the Methodology of Science

All of this is interesting and fun perhaps, but the mass of the day-to-day use of statistics is in the sciences—physical, life, and social—and in engineering. We will discuss the science role of statistics in some detail later in the chapter, but at this stage the question arises whether a nonscientist needs to have any knowledge of these questions at all. If you are not a scientist of any type and have no interest in the subject at all, why should you worry about ostensible scientific debates in general and those about health and environmental issues in particular? Why should you concern yourself with the essential role that statistics plays in the debate?

We have already answered this; because the decisions that are being made will affect you and your children directly and indirectly. The decisions made for good or bad reasons, with knowledge of the facts or not, and with or without reasonable interpretations of the data, will affect your income, lifestyle, and freedom to choose, not to mention your health. Perhaps some knowledge of how statistics in the sciences are used and, much more important, abused in the evaluation of data will help you in your evaluation of the research that will alter your life. The interpretation and the misinterpretation of statistics are being used every day in ways that affect almost all aspects of your life, economic, political, and social. One should think this is reason enough to study statistics and probability.

The study of statistics is itself a science. But first, let us briefly and intuitively consider the role of statistics in the sciences. Sciences are unified by the scientific method; they are distinguished by their subject matter. Statistics is a science, but as we will see it plays a unique role and has a unique subject matter.

The unifying scientific method is best characterized as an approach to understanding data rather than as a formal universal procedure that is to be rigidly followed at all costs. Even the physical sciences are not that tidy. The basic concept is the continuous interplay of hypothesis, observation, and measurement. In the early stages of a science, observation and measurement dominate; theorizing comes later.

The first step in this process is to create categories of similar objects, so that one can compare and contrast the objects within similar groups and between different

groups. Many of the life and social sciences are only just beginning to evolve from this stage. This part of the process involves measurement and the discovery of what characteristics to measure. The theory and practice of statistics is important even at this stage, because the central task is collecting and ordering information, or “statistics,” relevant to the science’s subject matter.

The second stage involves the proposal of hypotheses to explain whatever relationships or regularities are observed in the data. This is the stage involving the most use of imagination and insight and prepares the way for the third stage. The objective in this second stage of the development of the science is to create a story that provides a unifying framework into which most of the observations can be placed. More precisely, the objective is to provide a framework that allows us to interpret the observed relationships.

There are many ways, for example, to interpret the observed orbits of the stars and planets. We may invent a story about the Sun and the stars orbiting Earth as in the Greek myths or as in the religion of the ancient Britons who built Stonehenge in England, or we may postulate that the planets including Earth circle the Sun, or that planetary and Earth’s orbits around the Sun are ellipses. Each of these stories purports to provide a unifying idea that links all the observed relationships.

This second stage involves “modeling” the data; **modeling** is the process of creating a hypothesis that purports to explain the data. Essentially, in creating a hypothesis we build a model of the presumed mechanism that is generating the observations. *Explanation* in this context means that the hypothesis provides a set of logical links between groups of observations. This process enables us to say that if we observe one set of observations, here is the reason why we will see a second set of observations as a consequence. Both the theories of Copernicus and Kepler explained planetary and star observations and provided links between observations that enabled early astronomers to forecast even such rare events as eclipses.

As another example, the theory of demand in the study of economics provides a reason why we should expect to see a decrease in the quantity of a product consumed when the price is raised. In this example, the two groups of observations are “quantity consumed of a good” and “price.” The reason for the link between the two sets of observations is provided by the theory of demand based on the concept of maximization of individual utility. The process of specifying the causal link is the modeling of the data.

We now come to the crux of the scientific method: how to choose between competing hypotheses. The time-honored method is to choose the story that best matches the data, or provides “the best fit to the data.” This is the third stage in the development of a science and the best known: the “testing of hypotheses” stage.

We now face two difficulties. The first is that none of the hypotheses fits any of the data precisely; there are known errors of observation, not to mention the fact that the hypotheses are at best idealized abstractions from reality. The second is that we have to decide how to determine which of the alternative stories provides the best fit to the data; indeed, we even have to decide what we mean by “the best fit.” This is where the study of statistics and probability provides the whole of science with the basic methodology needed to answer these pervasive and vital questions. The development of these procedures is the central aim of this book.

In the previous few paragraphs we stated that the language of statistics provides the methodological tools that are needed by all the other sciences to evaluate their

theories and to measure the advance of knowledge in the field. The process of rejecting some hypotheses against others requires both the language and the techniques of statistics to be implemented. It is not too strong a statement to make that the other sciences could not advance without statistical theory and practice. The language of statistics is said to provide a *metalanguage* for the evaluation of the methodology used in the sciences; a **metalanguage** is a language that is independent of the language used to formulate the theories of the discipline under review. The applicability of the procedures generated by the field of statistics does not depend on the correctness, or “truth-falsity,” of the hypotheses under examination. This is a most useful property. If the validity of our testing procedures depended on the truth of the very hypothesis that we were trying to check, the results of any such test would be useless.

1.5 Statistics as a Science

The discipline of statistics is unique. Although the language of statistics provides a metalanguage for the analysis of any other discipline’s methodology, there is no other language in which to analyze the methodology of statistics; it cannot evaluate itself. The sole criterion for the success of statistical theory is the pragmatic one of whether or not the theory works. The evidence so far is that it does.

But the study of statistics is unique in another respect. We have said that sciences are distinguished by their subject matter. The subject matter of physics is in the physical properties of matter and energy; the subject matter of chemistry is in the chemical properties; that of economics is in the class of economic properties—that is, those having to do with the value of goods in exchange; that of biology is in the physical and chemical processes involved in living organisms; and so on.

The Subject Matter of Statistics

But what is the subject of statistics? In one sense, it is the subject matter of all the other disciplines; and in another sense, it is none of these. The subject matter of statistics is randomness, or stochastic variation. But randomness may be present, indeed is likely to be present, in the observation of all the data of all the other disciplines. In this sense, the subject matter of statistics includes all types of data. The subject matter of statistics is unique in that randomness is a property peculiar to the study of statistics. Let us briefly and intuitively examine this notion of randomness and see how it distinguishes statistics from all other disciplines. The terms to be used below will be explained more fully in the next chapters, but for now let us try to get a basic understanding of the main ideas.

The explanations, or stories, that we hypothesized about the relationships that exist in other disciplines are mainly *causal*. By this we mean that the story, or explanation, purports to show how the variation in one set of variables affects the values taken by another set of variables. Newton used his theory of gravity to explain the orbits of the planets; Boyle defined the relationship between temperature, pressure, and the volume of a gas; and demand theory in economics relates levels of consumption to levels of income. The variables in each of these examples are, respectively, *planetary orbits*; *temperature*, *pressure*, and *volume*; and *consumption* and *income*. In all these examples, the relationships are expressed functionally; that is, the relationship is an “if, then” type. The values of one set of variables are being determined by, at least according to the theory, the other set of variables.

In the simplest of cases one variable is controlled and so is called **deterministic**. If the other variable is functionally related to our deterministic, or controlled, variable, then it is also called deterministic. If I pluck a taut string of fixed length and mass with a given degree of tension, I will produce a specific tone, or frequency of vibration. The tone is functionally determined by the circumstances of the experiment, such as the nature of the material of the string, the degree of tension, and the length of the string. If we alter the length of the string, which is a controlled variable and so is deterministic itself, the theory of harmonic oscillators provides a function that predicts the tone that will be heard; so the tone is by our definition a deterministic variable as well.

The variables that are the subject matter of statistics are not deterministic; they are said to be random. A naive definition of a **random variable** that will do for now is one for which there is no known causal explanation. Random variables are not predictable. If we had an explanation for the variable, we would use it and there would be no need for a concept of randomness. Random variables are variables that are not part of a simple “if, then” type of explanation. These really are a different type of data. You are familiar with examples of such data: the individual outcomes of the tossing of coins or dice, drawing cards from a well-shuffled deck, roulette numbers, lotteries, the timing of lightbulb burnouts, the sex of offspring, tire blowouts, the paths of gas molecules, and so on. What is common about all these examples is that, given our current level of understanding, we cannot predict individual outcomes. We cannot provide a causal explanation for the occurrences of any of these phenomena as we have discussed them. We cannot specify through a theory a functional relationship between some other deterministic variable, or some variable that we control, and that we are seeking to explain. Statistics is all about *acausal* data. This explanation is simplified, of course, to stress the distinction between the two types of variables, or data, that the sciences use.

If statistics is the science of variables with no explanation, no theory of causality, then we have to wonder what might be meant by “explanation” in the context of such data. We might also question the earlier claim that in one sense statistics has as subject matter all of science.

Let us tackle the “explanation” question first. Clearly, the explanation provided by statistics is not going to be the same as that used in the other sciences. We will spend a few chapters on this very topic; you cannot expect to get a full idea within a few paragraphs. The explanation provided by the theory of statistics is descriptive of properties of large collections of such data; that is, we eschew trying to provide specific statements about individual occurrences and content ourselves with statements about large collections of observations. The models in such a theory are called **distributions**, and they are used to characterize the statistical properties of large collections of the data. What sort of properties these are we have yet to discover. For the moment, some brief ideas that we could describe include where the data seem to be centered, to what extent they vary about the center, whether there is symmetry, whether there is a small number of very large observations, and so on.

Statistics and Science Intertwined

The link to the other sciences is now not as difficult as it appeared to be just a few paragraphs ago. We have already mentioned that real data—the data that are actually observed in experiments or in the real-time functioning of an economy, an ecological

system, or the orbits of the planets and the stars—are all observed with error. These are errors of observation, errors of measurement, errors of recording, or errors of control over the external circumstances that can impinge upon the mechanism under inspection. This last example is very important in that our control over our experiment, or rather our lack of control, most often leads to false conclusions. Lack of control is exemplified by experiments that do not adequately allow for contamination of the results, by surveys that contain hidden biases to the responses, by collecting incomplete data on the experiment, and so on. Almost by their very nature, we can claim that such errors will have no causal explanation that would allow us to extend our “causal theory” to causal explanations of error. If there were such an explanation, then we could incorporate it into our theory; but experience teaches us that we will end up with an irreducible residue of acausal, or random, effects.

We can guess that for large numbers of measurements that the average error might not be so large. Almost instinctively, we all do this in our daily lives. Whenever the length is critical, we measure the curtains several times; we may not know where any specific tennis serve is going to land, but on average, we have a very good idea; our trip to the university never takes the same time, but on average we know what to do to avoid being late.

Even in the simplest of situations, observed data in the physical, biological, and social sciences contain two components (at least in theory): a deterministic part that is described directly by the theory and a random component that must be described by statistical reasoning. Consequently, the empirical validation of all sciences and the refutation of hypotheses requires a methodology that blends deterministically formulated relationships with a random component. This is in fact what happens in all science. The difficulty is to separate the two components and to discover how to “remove the stochastic veil,” or the veil of randomness, to discover the actual relationship that is buried in *noise*, as the random variation is often called. Sometimes the level of noise is so great that the underlying signal is impossible to detect without very sophisticated tools, and often not even then. Atmospheric data are very noisy, which is why, in part, trying to detect small, long-run changes is so very difficult; such efforts are a real test of the power of statistical procedures.

If this were all there were to the matter, the science of statistics would be useful, but not very interesting in its own right. As it turns out, there are relationships in science that can only usefully be formulated in terms of random variables. We previously stated that the theory of statistics is embedded in the distributions of collections of data, so the explanations provided by statistical reasoning are in terms of the properties of these distributions. Many scientific relationships must be couched in terms of the distributions of the variables rather than in terms of the variables themselves. This approach is most clearly exemplified in terms of thermodynamics, in describing the dynamical processes involved in chemical reactions, and most likely in the whole of the social sciences. We do not try to predict what a particular individual will do in responding to a price change but what will happen with a large collection of people all facing the same circumstances. We do not try to predict where an individual gas molecule will be within a container but content ourselves with a macrodescription of an enormous collection of molecules, such as specifying the pressure of the gas at a given temperature.

The difference in the way theory is formulated with respect to “distributions, or large collections, of random variables” and the usual formulation in terms of “if, then” relationships is a very important one. The distinction changes the way in which a scientist

views explanation and certainly changes the way in which one decides how to choose between alternative models of behavior. As we will see, a distribution is a way of describing collections of events and is itself modeled by a mathematical function. The change in viewpoint can now be characterized by a shift from simple causal relationships between deterministic variables to a relationship between a variable, random or deterministic, and a function; as the variable changes, so does the function that models the distribution of the other variable. If the size of a container of a gas is suddenly expanded, for example, the distribution of the gas molecules will shift over time in reaction to the new conditions; that is, we have to change our description of the collection of positions of the molecules. If the price of a commodity is lowered, the distribution of levels of consumption of that good by a group of consumers will shift to reflect the new circumstance; that is, our description of the collection of individual levels of consumption will change. If there is an increase in the demand for a product, the distribution of costs across firms in the industry will shift to reflect the change; or we restate the matter by saying that our description of the collection of costs changes in reaction to the change in demand.

We now see that an important contribution of statistical thinking is in the very formulation of some of the theories in all the other sciences. The introduction of random variables into scientific models shifts the concentration from trying to describe the specific paths of individual elements to the behavior of large collections, or “ensembles,” of data and from simple “if, then” relationships to explaining the factors that change distributions of random variables. We now see that the theory of statistics and probability is inextricably linked to all the sciences. Not only does the study of statistics provide the methodology for the empirical evaluation of science, but it provides the basic concepts for the conceptualization of many of the relationships in the sciences.

Whatever you choose to study in science, whatever science that you need to use, whatever policy that requires the physical or social sciences as basic information, that choice will necessitate an understanding of the theory and practice of statistics. Now you understand why statistics is a part of almost any university curriculum, especially in the sciences (physical, life, and social as well as in engineering).

1.6 Summary

In this first chapter, we have introduced several new ideas and reviewed a few others.

Our first lesson is that the discipline of statistics is concerned with the problems generated by decision making and the need to evaluate data for either scientific or policy reasons. We presented examples to indicate the wide range of alternative situations in which a statistical approach was needed to tackle the problem. We claimed that the discipline of statistics provides the framework for analyzing decisions, although the precise formulation of such a framework is to be left to later chapters.

The case was made that statistics is a science in that it shares with all science the same methodology. Actually, the theory of statistics provides the methodology for evaluating data and for the testing of hypotheses that is used by all the sciences. The discipline of statistics involves the process of choosing between alternative explanations of observed data. In this regard statistics is unique.

The subject matter of statistics, narrowly defined, is also unique. The subject matter is random, or “acausal” data—that is, data for which there is no known causal explanation or prediction. Whereas the simplest theoretical models in the sciences

involve “if, then” relationships between the observed values of the science’s variables, the situation in statistics is more complex.

Because statistics deals with data for which there is no causal relationship, “explanation” in statistics means specifying a distribution for large collections, or ensembles, of random variables; such distributions become the models for an acausal science. We give up trying to predict the specific values of a particular observation for saying something about the properties of large collections of observations.

The connection with the sciences became clear when we first realized that all variables observed in science involve errors of observation, and that these errors are essentially acausal in nature. The empirical verification of any science involves statistics because the observed variables have two components: one that might be deterministic and one, due to error, that is random.

The next step in the development of our understanding led to the realization that the concepts of statistics often are involved in science at a more fundamental level. Some parts of the physical sciences—such as thermodynamics, and probably the whole of the social sciences—are best formulated as relationships between variables and those functions that we have called “distributions.” This insight involves a fundamental shift in thinking about the meaning of both causality and the concept of explanation in science. In the expanded view, causality must be expressed in terms of shifts in distribution functions—that is, in terms of descriptions of collections of events rather than of naive “if, then” relationships between the levels of the variables of the system.

The next chapter will elaborate on these topics at greater length and clarify some of the terms that we are beginning to use.

Case Study

Was There Age Discrimination in a Public Utility?

A small public utility was facing the impact of new organizational structures within its industry. This necessitated regrouping production facilities and redetermining the use of its labor force to obtain the necessary gains in efficiency that would enable it to survive. An outside consulting firm was called in to advise it on setting up new procedures and to design new job descriptions.

The restructured firm’s procedure to deal with this situation was to have everyone resign from the “old” firm and then reapply to the “new” firm along with other potential employees recruited through advertising for the “new” jobs with the new job descriptions. This change was imposed on all em-

ployees, including senior management. New employees were hired from the enlarged pool of applicants in three stages; the most senior managers were hired first, then intermediate level managers, then everyone else.

The union representing the old, lower-level employees promptly sued the new firm for age discrimination, claiming that the distribution of new hires in the new firm had a much younger average age of employees than was true for the old firm. Statistically trained economists were hired by both parties to aid the lawyers in resolving this problem.

In this text, we will explore this issue in depth using the tools that we shall develop. Our first task is to note the role played by statistical analysis in this court case. The amount of money involved was approximately \$42 million and the number of

continues on next page

(Continued)

ex-employees involved in the suit was 52, so the issues involved were substantial.

Here are some of the questions that we might ask to deal with this issue. We might compare the age of those hired with those not hired and with the age of the original employees. How does the age of those employees rehired compare with that of the external candidates that were hired? How do the salaries before and after the restructuring compare, especially for those over 40 years of age, and how does this difference, if any, depend on the breakdown between hired and not hired, and internal versus external, candidates? The real issue is whether there was discrimination by age, and did that discrimination have an impact on the wages paid. But to answer these questions we will have to separate the effects of other factors that may be related to age and salary levels coincidentally.

We will carry this case forward through all the chapters of the book to show you how the analysis that you will learn applies to this real-life situation. We seek to show you how the study of statistics will guide you in recognizing the important is-

ssues, how to determine a procedure for analyzing the data, and how to decide what data to collect and in what manner. Finally, we hope to be able to demonstrate both the strengths and limitations of statistical analysis in this and similar situations. But mainly, as you progress through each chapter, you will discover how much you will learn as you master more and more statistical analysis.

You should also recognize that the data set listed in the Xfiles is a real data set used in an actual court case. As such, there are inevitable errors, misstatements, and inconsistencies; look out for these difficulties. Being on guard for procedural and recording errors is also part of learning to be a statistician!

The key to what follows is not that “as a statistician” you will have more data available, but that you will be able to extract far more information from the data that you do have. The primary gain for decision making that a statistician contributes is his or her interpretation of the existing data—that is, the ability to determine the extent of the relevance, or the lack of relevance, of the available data to the solution of the problem at hand.

1.7 Addendum for the Reader

You will soon discover, if you have not done so already, that this text differs from others. This section indicates some of those ways and outlines the implications for how you read the book and address some of the exercises.

You may have already realized that in this text we focus much more on the understanding of *processes*, of discovering new ways of viewing the world, not just learning new terms and symbols. Together, we are trying to tackle a more difficult task than you may be used to considering: the discovery of a new way of thinking about the facts that you observe, a new way to organize data. Such a process is more challenging and requires more effort on your part than the simple acquisition of a sequence of facts. Although this might all sound obvious, it makes a considerable difference in the way you approach the text and particularly in how you interpret some of the questions in the exercises. Let me explain.

First, let us review the best way to read the text. I suggest that for each chapter, you begin by looking at the title, read the “What You Will Learn in This Chapter” section, look through the Introduction, note the section headings, and read the Summary.

This will give you an overview of what is in the chapter and some idea of the terms that will be introduced. It also will provide a rough goal to be achieved in mastering the chapter. This step will usually take only a few minutes and is soon well rewarded.

The second stage is to read through the whole chapter quite quickly. Your task at this stage is to note the sections that you find fairly easy to understand, note those where you have difficulties, and determine the terms and formulae that you will have to learn and remember. Most important, the second reading is to enable you to acquire a broad overview of the material in the text. Do not at this stage try to slog doggedly through each paragraph, one by one, mastering every detail as soon as it is met. Do pay attention to the graphs; switch back and forth between the texts and the graphs, integrating the two, and observe how the graphs illustrate and illuminate the text.

The third stage is to review the whole chapter, trying first to be sure that you do understand the easier sections. You can do this by querying yourself about what the text means and by summarizing the argument in your own words. More important, when you come to the difficult sections, you need to ask yourself some very important questions concerning your difficulties. Try to figure out *why* you are having trouble.

The simplest reason may be because the text is using notation, concepts, or mathematical tools with which you are unfamiliar. Before proceeding any further, make sure that you understand the terms and concepts that the text presumes you know. These may be ideas introduced earlier in the text that you did not master at that time, so go back and make sure of the earlier material. The difficult concepts may be mathematical ideas that you have forgotten, so review the mathematical appendix. Or you may decide that the idea being presented is one that you just do not understand.

If the last case is the situation, concentrate on the difficult passage. Reread it carefully, line by line, querying yourself about your understanding at each step. Try to formulate your own simple examples that attempt to illustrate the ideas being discussed. Look at the exercises that illustrate the idea, especially the worked exercises. Best, try to draw a picture of what is going on, and write down your own efforts to present the argument. If you like, pretend that you are trying to explain this concept to someone else; often the very attempt to do that will suddenly reveal to you what is involved. Talk to other students about the problem; what may be difficult for you in one situation may be easy for someone else. Later, you will be able to return the favor in a passage that you find easy and they find difficult. Remember that you are trying to understand a “process,” the way in which statisticians view the world. This requires you to develop new skills and new ways of thinking. It is inevitable, if you are learning new ways of thinking, that there will be times when you will see the logic of the process only after some effort. If you have ever learned a new sport, say tennis, you will recognize that it took some time to learn how to process information on the ball’s flight, the properties of the racket, and the mechanisms that you needed to develop to be able to control your strokes. Similarly, in this situation you are trying to discipline your brain to think in certain ways that may well be very novel to you. This will require practice and analysis of your mistakes.

Having worked your way through the chapter, reread it and write down your own summary of the chapter. Include all formulae that you may need to remember. Your notes should be sufficiently complete to enable you to answer any question about the material.

With respect to the problem of remembering formulae, note that the main task in the text is to *understand* the equations and formulae; the text is written to facilitate exactly

that process. However, while you are learning, and especially during examinations, it is more efficient if you can remember the various formulae and equations without having to look them up every time. There is a painless way to master the required memorization. Obtain a set of $3'' \times 5''$ index cards. Every time you come across a formula, a definition, or an equation that you want to remember, write it down on the card; limit it to one per card. Carry these cards with you always. Then, when you are waiting for a bus, a plane, or a friend, or you are standing in line at the supermarket checkout counter, take out a few of these cards and refresh your memory. See how much you can remember without looking and then review the card. Very soon and with seemingly little effort, you will remember more concepts than you ever thought you could manage.

The exercises are somewhat different from what you may have experienced before. Beginning with Chapter 2, exercises in each chapter are divided into three sections. The ones in the first section, "Calculation Practice," will be fairly simple and are the closest to your normal experience. Even here, recognize that you are developing a new skill, the ability to perform calculations easily and without hesitation. The objective here is for you to reach a state in which you no longer have to think about the mechanics of your calculations. The skills that you will develop in this first section proceed in stages. The first stage will use very simple numbers to make you familiar with the calculations. The second stage introduces the computer as a device to relieve you of the tedium of adding up products of numbers. For more complicated formulae, after you have gained experience in pulling together all the components by hand, you will be able to use the computer to perform all the calculations. Do this only after having verified that the computer algorithm is in fact producing the same results that you have obtained for yourself.

The exercises in the "Exploring the Tools" sections do require you to think a lot more about what you are doing and why. You are trying to explore the meaning and proper use of the tools by considering their use and properties in unusual circumstances; you are being challenged to think about what you are doing, not just to perform repetitious calculations. Many of the questions require you to make a series of similar calculations and then ask you to compare your results. This is a request for you to think about your results to try to gain a better appreciation for what you are doing. Requiring you to consider the similarities and differences between a series of related calculations facilitates this process. When you are requested to compare results, look for the way in which the results are similar, or different, and how those similarities or differences vary as you vary the parameters under your control.

The exercises in the "Applications" sections require the most thinking. You will soon learn that the major question in Applications is, "What is the statistical question?" and the most important outcome from the analysis is, "What did I learn from the exercise?" Consequently, many of the questions in the Applications sections will be open ended with no simple formulaic response. Many of the questions ask for your evaluation of the results that you obtain. The more you learn and the better you understand statistical theory, the more sophisticated and useful will be your insights and interpretations of the data you observe. Indeed, the entire objective of any statistical analysis is to provide useful interpretations of the data. Many questions in the exercises ask you to compare results of two different calculations, or results of the same calculation on different data. Your answers here require your use of judgment, insight into the nature of the problem, and an ability to determine how to use the tools to address the essential issues involved. Indeed, determining the essential issues is itself a

key part of the question. These questions are a far cry from the type of question exemplified by:

If a train is traveling on a track at 50 miles per hour and left its station at 10:00 and another train left its station 300 miles away at 10:00 and is traveling at 40 miles per hour in the opposite direction, at what time will the trains meet?

There is no room here for having to determine what the question is, nor for deciding how to solve it, nor for deciding what data are needed to answer the question, nor for interpreting your answer. One merely carries out a calculation and quotes an answer that is either right or wrong. The problems that we begin to address in the Applications sections are not so simple. You will need to decide what the relevant question is, how you are going to solve it, what data you need, what procedures you should use, and how you interpret the results. But of course, this reflects the questions that you will have to deal with in real life.

In answering the questions in the exercises, especially in the Exploring the Tools and Applications sections, you will often have to return to the text to review what is written there and to reconsider what you thought you understood before you began to answer the question. Do not be discouraged by this process of repeatedly reviewing the text; this is part of the learning process. We only really learn when we are confronted with having to do something, to make a decision. It is then that we discover that what we thought we knew, we do not. This is the most important objective of the exercises, so do expect to reread the text along with the graphs and the exercises.

There are many “worked” exercises. These are examples of how to use statistical thinking to analyze a problem or guidance in how to learn more about the tools that you are trying to acquire. They also provide a blueprint for doing similar exercises on your own. These exercises, which are always introduced with a sentence that gives the objective of the exercise, are most important. If you do no other exercises, at least do these. If you are unsure how to answer a question that you have had assigned, look back to a worked exercise that is close to the question you are trying to answer. Work through the worked exercise, and then you should have no difficulty in doing the assigned exercise on your own.

I hope that the computer exercises will encourage you to explore the tools and concepts you are trying to master. Feel free to try anything you fancy in these exercises, especially anything that I have not already suggested. Be bold; you cannot damage anything.

I trust that you will enjoy reading the text and learning the basics of statistics from it. A mastery of elementary statistical reasoning is a major component of your training and one that you will use in all aspects of your life; so it is comforting if you can enjoy the process as well. No matter what you do in life, you will need to evaluate statistics and statistical reasoning on almost a daily basis; so a good working knowledge of statistical processes is important for success in your chosen career.

Exercises

Topics for Consideration

Before beginning the exercises, you should read the “Addendum to the Reader” at the end of Chapter 1 if you have not already done so.

It would be advantageous for you to keep your answers to these questions on file. In Chapter 15, you will be asked to reconsider these questions in light of the analysis that you will learn in the chapters to follow.

1.1 According to a *Wall Street Journal* (March 28, 2000) article:

Most years, a majority of portfolio managers trail the market. A big reason is the drag from investment costs, including management fees and trading. But the pattern of stock returns also plays a role. The market average (which is a weighted mean) tends to be driven by a fistful of stocks that post huge gains, so that the median stock lags behind the market. Result? Each year, you end up with a minority of money managers who hold the year's hottest stocks and thus earn fabulous returns, while the rest miss out on the big winners and therefore trail the index.

Note the terms that you need to know to be able to follow the argument, *mean* and *median*, for example. Discuss this quote. Try to address the following issues:

- What data would you collect to confirm or deny this claim?
- What would constitute rejection of the author's claims, and how would you assess the matter?
- Assess the impact of these statements on your own choice of investment portfolio.

1.2 According to a *Wall Street Journal* (January 18, 2000) article:

Kyoto [conference on global warming] supporters cheer new findings that the Earth's surface temperature is probably rising. But this trend isn't recent and isn't man made. (Arthur and Noah Robinson)

The authors point out that the current warming trend began about 300 years ago after the bottom of the Little Ice Age and that current temperatures are far below those of about a thousand years ago. Further, an average increase in temperature will increase the growing seasons and extend their geographical limits. *Scientific American* (August 2000) counters that diseases common to tropical climates will spread north.

Debate how you would evaluate these claims. Explore the need for data, reliable models of the environment, methods for determining the accuracy of

the model's forecasts, and how one would balance increases in farming productivity with increases in tropical diseases. In the process, you should recognize the need for the analysis of data subject to much uncertainty. Demonstrate the manner in which the scientific method is relevant to this analysis.

1.3 According to a *New York Times* (October 31, 1999) article:

A ranking is easier to convey information to a large public, like a football ranking. . . . We can clearly say that Denmark is cleaner than Germany and Germany is cleaner than Botswana. But in between there may be some problems—saying that Kenya is cleaner than Tanzania, for example.

The article goes on to report that Bolivia's presidential election in 1997 was won on the basis of an anticorruption campaign that was bolstered by quoting rankings showing that Bolivia had a worse ranking on corruption than its neighbors.

Indicate the dangers of uncritical reliance on such figures. Debate an alternative strategy for evaluating the relative nature of corruption in two countries. What are some of the difficulties in collecting the appropriate data?

1.4 "Obesity Rate Rising Fastest in the South" was a headline in the *New York Times* (October 27, 1999) based on a telephone survey of 100,000 people in 1991 and 1998. The two samples used different people. Explore how you would evaluate the quality of the data that were used to make the claim that southerners' obesity rate jumped 67.2%.

1.5 In the *Economist* (August 5, 2000, p. 61), there is a quote from the American Consumer Satisfaction Index that a University of Michigan Business School indicator of consumer satisfaction shows that consumer satisfaction has fallen substantially in almost all areas since 1994 but mainly in the service industries. Discuss how you would devise a strategy to determine whether the decline represents more particular customers who are more willing to complain or a real decline in the services provided by firms.

1.6 “Scientist Differs with EPA on Malathion” is a headline in the the *New York Times* (August 18, 2000). The article indicates that Dr. Brian Dementi claims that the use of malathion to control mosquitos is a greater cancer risk than the EPA admitted. Explore the extent to which this debate rests, or should rest, on the accuracy and relevance of data; the degree of confidence in the theoretical science; and the procedures used to evaluate the data. Do you perceive how objective scientists might well differ on their recommendations whether or not to use malathion, even though they both have access to the same experimental and survey data?

1.7 In August 2000, Ford and Firestone had to recall millions of Firestone tires that had been declared potentially unsafe. Imagine that you are a manager at Firestone assigned the task of trying to assess the cost of this recall. Discuss the data requirements and what sort of modeling will be needed to come up with a “reasonable figure.”

1.8 In the *Economist* (August 19, 2000), an article “Asbestos Claims Still Killing,” points out that the claims against asbestos manufacturers are not only huge but growing rapidly. Indeed, the outstanding claims already are greater than the total net worth of the firms involved. Explain how you would use and assess the value of information in analyzing this situation.

The original research on the carcinogenic effects of asbestos were limited to specific types of asbestos and to workers in asbestos plants. Discuss the relevance of this evidence to situations in which asbestos of different types to those examined are installed and are not disturbed or “worked.”

1.9 A quote from the *Economist* (August 5, 2000, p. 58) follows:

Even in good times, some 40% of all new businesses fail within their first five years of operation, according to Census Bureau figures. In America as a whole,

more than 70,000 companies went bust during 1998, almost half as many as were started that year.

By contrast, the drumbeat of Internet failures is still incredibly faint. There are at least 3,000 dot.com start-ups backed by American venture-capital firms, according to VentureOne, a consultancy. . . . Dotcomfailures.com lists a mere 20 [failures].

Imagine that you are analyzing investments in e-commerce companies. Explain how you would bring data and modeling to bear on this problem in trying to decide the near-term profitability of e-commerce firms.

1.10 The *Scientific American* (September 2000) included an article concerning the attempt to determine the relative importance of genes versus training for championship athletes. One finding is that sprinters and marathoners have very different proportions of slow versus fast myosin filament in the muscle fiber. Consider the role of data, modeling, and analysis in determining these results. How much credence do you suspect they have? Discuss the role of the scientific method in this process.

1.11 Airbus has committed itself to building the world’s largest passenger airplane. Imagine that you are the engineer in charge of determining the potential cost to build and operate this plane. Alternatively, you might be on the marketing side and need to assess the financial viability of such a venture in the market. Discuss the strategy that you might follow to fulfill your task. Evaluate the data needs and how you would assess their accuracy and relevance.

1.12 “Worrying about Wireless. Researchers are still unsure whether cellular phones are safe” reads a headline in the *Scientific American* (September 2000). Explain how the resolution of this issue is one of relating data to theoretical concepts and being able to assess the accuracy and relevance of the data collected. Then discuss issues of how to collect the data.