INTRODUCTION

Uses of Probability and Statistics

Introduction

This chapter introduces you to probability and statistics. First come examples of the kinds of practical problems that this knowledge can solve for us. Next this Introduction discusses the relationship of probabilities to decisions. Then comes a discussion of the two general types of statistics, descriptive and inferential. Following this is a discussion of the limitations of probability and statistics. And last is a brief history of statistics. Most important, the chapter describes the types of problems the book will tackle.

Because the term "statistic" often scares and confuses people—and indeed, the term has several sorts of meanings—the chapter includes a short section on "Types of Statistics." Descriptive statistics are numbers that summarize the information contained in a group of data. Inferential statistics are procedures to estimate unknown quantities; that is, these procedures infer estimates and conclusions based on whatever descriptive statistics are available.

At the foundation of sound decision-making lies the ability to make accurate estimates of the probabilities of future events. Probabilistic problems confront everyone—from the business person considering plant expansion, to the scientist testing a new wonder drug, to the individual deciding whether to carry an umbrella to work.

To those interested only in learning resampling statistics, or who have some previous acquaintance with probability and

statistics, I suggest that you merely glance over this Introduction and then proceed directly with Chapter 1.

If you want to find out quickly about the resampling method, you might skim the next few chapters and perhaps skip all the way to Chapter 15.

What kinds of problems shall we solve?

These are some examples of the kinds of problems that we can handle with the methods described in this book:

1. *You are a doctor* trying to develop a cure for cancer. Currently you are working on a medicine labeled CCC. You have data from patients to whom medicine CCC was given. You want to judge on the basis of those results whether CCC really cures cancer or whether it is no better than a sugar pill.

2. *You are the campaign manager* for the Republicrat candidate for President of the United States. You have the results from a recent poll taken in New Hampshire. You want to know the chance that your candidate would win in New Hampshire if the election were held today.

3. *You are the manager and part owner* of a small construction company. You own 20 earthmoving trucks. The chance that any one truck will break down on any given day is about one in ten. You want to know the chance on a particular day—tomorrow—that four or more of them will be out of action.

4. *A machine gauged to produce screws* 1.000 inches long produces a batch on Tuesday that averaged 1.010 inches. Given the record of screws produced by this machine over the past month, we want to know whether something about the machine has changed, or whether this unusual batch has occurred just by chance.

The core of all these problems, and of the others that we will deal with in this book, is that you want to know the "chance" or "probability"—different words for the same idea—that some event will or will not happen, or that something is true or false. To put it another way, we want to answer questions about "What is the probability that...?", given the body of information that you have in hand.

The question "What is the probability that...?" is usually not the ultimate question that interests us at a given moment.

Eventually, a person wants to use the estimated probability to help make a *decision* concerning some action one might take. These are the kinds of decisions, related to the questions about probability stated above, that ultimately we would like to make:

1. *Should you (the researcher) advise doctors to prescribe medicine CCC* for patients, or, should you (the researcher) continue to study CCC before releasing it for use? A related matter: should you and other research workers feel sufficiently encouraged by the results of medicine CCC so that you should continue research in this general direction rather than turning to some other promising line of research? These are just two of the possible decisions that might be influenced by the answer to the question about the probability that medicine CCC cures cancer.

2. *Should you advise the Republicrat presidential candidate to go to New Hampshire* to campaign? If the poll tells you conclusively that he or she will not win in New Hampshire, you might decide that it is not worthwhile investing effort to campaign there. Similarly, if the poll tells you conclusively that he or she surely will win in New Hampshire, you probably would not want to campaign further there. But if the poll is not conclusive in one direction or the other, you might choose to invest the effort to campaign in New Hampshire. Analysis of the chances of winning in New Hampshire based on the poll data can help you make this decision sensibly.

3. *Should your firm buy more trucks*? Clearly the answer to this question is affected by the probability that a given number of your trucks will be out of action on a given day. But of course this estimated probability will be only one part of the decision.

4. *Should we adjust the screw-making machine* after it produces the batch of screws averaging 1.010 inches? If its performance has not changed, and the unusual batch we observed was just the result of random variability, adjusting the machine could render it more likely to produce off-target screws in the future.

The kinds of questions to which we wish to find probabilistic and statistical answers may be found throughout the social, biological and physical sciences; in business; in politics; in engineering (concerning such spectacular projects as the flight to the moon); and in most other forms of human endeavor.

Probabilities and decisions

There are two differences between questions about probabilities and the ultimate decision problems:

1. Decision problems always involve *evaluation of the consequences*—that is, taking into account the benefits and the costs of the consequences—whereas pure questions about probabilities are estimated without evaluations of the consequences.

2. Decision problems often involve a *complex combination* of *sets of probabilities and consequences*, together with their evaluations. For example: In the case of the contractor's trucks, it is clear that there will be a monetary loss to the contractor if she makes a commitment to have 16 trucks on a job tomorrow and then cannot produce that many trucks. Furthermore, the contractor must take into account the further consequence that there *may* be a loss of goodwill for the future if she fails to meet her obligations tomorrow—and then again there *may not* be any such loss; and if there is such loss of goodwill it might be a loss worth $10,000 *or* $20,000 *or* $30,000. Here the decision problem involves not only the probability that there will be fewer than 16 trucks tomorrow but also the immediate monetary loss and the subsequent possible losses of goodwill, and the valuation of all these consequences. The complexity of the contractor's decision problem may be seen in the schematic diagram called a "decision tree" shown in Figure i-1.

Whether to buy more trucks

Continuing with the decision concerning whether to do more research on medicine CCC: If you do decide to continue research on CCC, (a) you may, or (b) you may not, come up with an important general cure within, say, the next 3 years. If you do come up with such a general cure, of course it will have very great social benefits. Furthermore, (c) if you decide not to do further research on CCC now, you can direct your time and that of other people to research in other directions, with some chance that the other research will produce a less-general but nevertheless useful cure for some relatively infrequent forms of cancer. Those three possibilities have different social benefits. The probability that medicine CCC really has some curative effect on cancer, as judged by your prior research, obviously will influence your decision on whether or not to

do more research on medicine CCC. But that judgment about the probability is only one part of the overall web of consequences and evaluations that must be taken into account when making your decision whether or not to do further research on medicine CCC. Again, the web of consequences and evaluations for the contractor deciding whether to buy more trucks is sketched in the decision tree in Figure i-1.

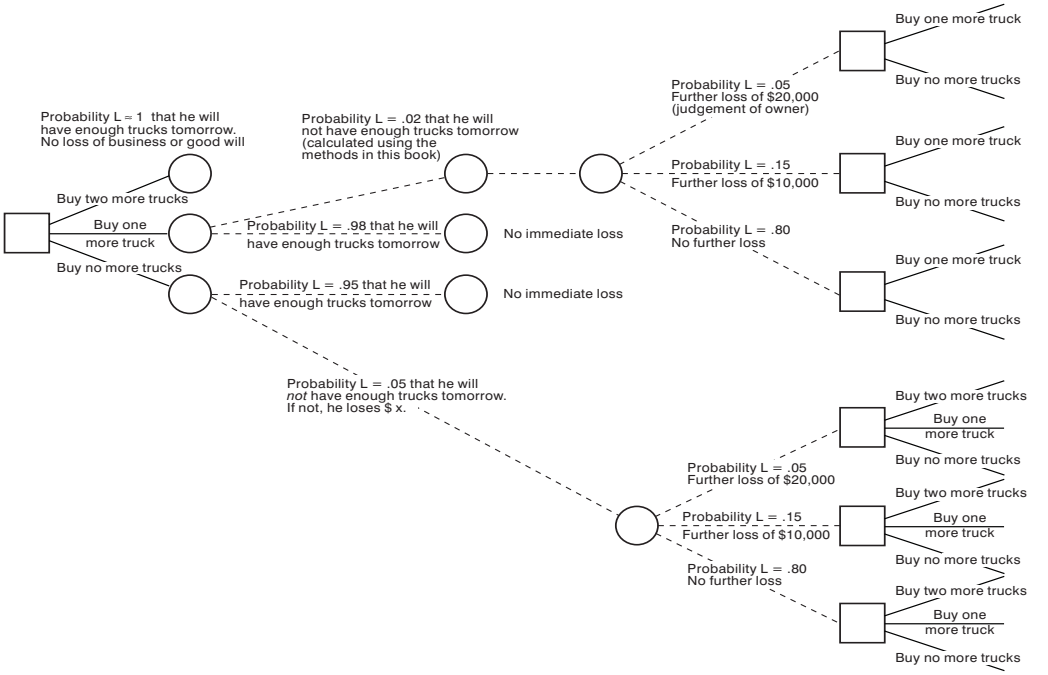

Figure i-1: Decision Tree

Why does this book limit itself to the specific probability questions when ultimately we are interested in decisions? A first reason is division of labor. The more general aspects of the decision-making process in the face of uncertainty are treated well in other books. This book's special contribution is its new approach to the crucial process of estimating the chances that an event will occur.

Second, the specific elements of the overall decision-making process taught in this book belong to the interrelated subjects of *probability theory* and *inferential statistics*. Though probabilistic and inferential-statistical theory ultimately is intended to be part of the general decision-making process, often only the estimation of probabilities is done systematically, and the rest of the decision-making process—for example, the decision

whether or not to proceed with further research on medicine CCC—is done in informal and unsystematic fashion. This is regrettable, but the fact that this is standard practice is an additional reason why the treatment of inferential statistics and probability in this book is sufficiently complete.

A third reason that this book covers only inferential statistics and not decision statistics is because most college and university statistics courses and books are limited to inferential statistics, especially those courses and books for students in the social sciences.

Types of statistics

The term *statistics* sometimes causes confusion and therefore needs explanation.

A statistic is a *number*. There are two kinds of statistics, summarization (descriptive) statistics and probability statistics. The most important summarization statistics are the total, averages such as the mean and median, the distribution, the range, and other measures of variation. Such statistics are nothing new to you; you have been using many of them all your life. Inferential statistics, which this book deals with, uses descriptive statistics as its input.

Inferential statistics can be used for two purposes: to aid scientific *understanding* by estimating the probability that a statement is true or not, and to aid in making *sound decisions* by estimating which alternative among a range of possibilities is most desirable.

Limitations of probability and statistics

Statistical testing is not equivalent to social-science research, and research is not the same as statistical testing. Rather, statistical inference is a handmaiden of research, often but not always necessary in the research process.

A working knowledge of the basic ideas of statistics, especially the elements of probability, is unsurpassed in its general value to everyone in a modern society. Statistics and probability help clarify one's thinking and improve one's capacity to deal with practical problems and to understand the world. To be efficient, a social scientist or decision-maker is almost certain to need statistics and probability.

On the other hand, important research and top-notch decision-making have been done by people with absolutely no formal knowledge of statistics. And a limited study of statistics sometimes befuddles students into thinking that statistical principles are guides to research design and analysis. This mistaken belief only inhibits the exercise of sound research thinking. Kinsey long ago put it this way:

However satisfactory the standard deviations may be, no statistical treatment can put validity into generalizations which are based on data that were not reasonably accurate and complete to begin with. It is unfortunate that academic departments so often offer courses on the statistical manipulation of human material to students who have little understanding of the problems involved in securing the original data. …When training in these things replaces or at least precedes some of the college courses on the mathematical treatment of data, we shall come nearer to having a science of human behavior. (Kinsey et al., 1948, p. 35).

In much—even most—research in social and physical sciences, statistical testing is not necessary. Where there are large differences between different sorts of circumstances—for example, if a new medicine cures 90 patients out of 100 and the old medicine cures only 10 patients out of 100—we do not need refined statistical tests to tell us whether or not the new medicine really has an effect. And the best research is that which shows large differences, because it is the large effects that matter. If the researcher finds that s/he must use refined statistical tests to reveal whether there are differences, this sometimes means that the differences do not matter much.

To repeat, then, some or even much research—especially in the physical and biological sciences—does not need the kind of statistical manipulation that will be described in this book. But most decision problems *do* need the kind of probabilistic and statistical input that is described in this book.

Another matter: If the raw data are of poor quality, probabilistic and statistical manipulation cannot be very useful. In the example of the contractor and her trucks, if the contractor's estimate that a given truck has a one-in-ten chance of being out-of-order on a given day is very inaccurate, then our calculation of the probability that four or more trucks will be out of order on a given day will not be helpful, and may be misleading. To put it another way, one cannot make bread without flour, yeast, and water. And good raw data are the flour, yeast and water necessary to get an accurate estimate of a probabil-

ity. The most refined statistical and probabilistic manipulations are useless if the input data are poor—the result of unrepresentative samples, uncontrolled experiments, inaccurate measurement, and the host of other ways that information-gathering can go wrong. (See Simon and Burstein, 1985, for a catalog of the obstacles to obtaining good data.) Therefore, we should constantly direct our attention to ensuring that the data upon which we base our calculations are the best it is possible to obtain.