CHAPTER 14

Framing Statistical Questions

Introduction

Chapters 3-10 discussed problems in probability theory. That is, we have been estimating the probability of a composite event *resulting from a system in which we know the probabilities of the simple events*—the "parameters" of the situation.

Then Chapters 11-13 discussed the underlying philosophy of statistical inference.

Now we turn to inferential-statistical problems. Up until now, we have been estimating the complex probabilities of *known* universes—the topic of *probability*. Now as we turn to problems in *statistics*, we seek to learn the characteristics of an unknown system—the basic probabilities of its simple events and parameters. (Here we note again, however, that in the process of dealing with them, all statistical-inferential problems eventually are converted into problems of pure probability). To assess the characteristics of the system in such problems, we employ the characteristics of the sample(s) that have been drawn from it.

For further discussion on the distinction between inferential statistics and probability theory, see Chapters 1-3.

This chapter begins the topic of *hypothesis testing*. The issue is: whether to adjudge that a particular sample (or samples) come(s) from a particular universe. A two-outcome yes-no universe is discussed first. Then we move on to "measured-data" universes, which are more complex than yes-no outcomes be-

cause the variables can take on many values, and because we ask somewhat more complex questions about the relationships of the samples to the universes. This topic is continued in subsequent chapters.

In a typical hypothesis-testing problem presented in this chapter, one sample of hospital patients is treated with a new drug and a second sample is not treated but rather given a "placebo." After obtaining results from the samples, the "null" or "test" or "benchmark" hypothesis would be that the resulting drug and placebo samples are drawn from the same universe. This device of the null hypothesis is the equivalent of stating that the drug had no effect on the patients. It is a special intellectual strategy developed to handle such statistical questions.

We start with the scientific question: Does the medicine have an effect? We then translate it into a testable statistical question: How likely is it that the sample means come from the same universe? This process of question-translation is the crucial step in hypothesis-testing and inferential statistics. The chapter then explains how to solve these problems using resampling methods after you have formulated the proper statistical question.

Though the examples in the chapter mostly focus on tests of hypotheses, the procedures also apply to confidence intervals, which will be discussed later.

Translating scientific questions into probabilistic and statistical questions

The first step in using probability and statistics is to translate the *scientific* question into a *statistical* question. Once you know exactly which prob-stats question you want to ask—that is, exactly which probability you want to determine—the rest of the work is relatively easy (though subtle). The stage at which you are most likely to make mistakes is in stating the question you want to answer in probabilistic terms.

Though this translation is difficult, it involves no mathematics. Rather, this step requires only hard thought. You cannot beg off by saying, "I have no brain for math!" The need is for a brain that will do clear thinking, rather than a brain especially talented in mathematics. A person who uses conventional methods can avoid this hard thinking by simply grabbing the formula for some test without understanding why s/he

chooses that test. But resampling pushes you to do this thinking explicitly.

This crucial process of translating from a pre-statistical question to a statistical question takes place in all statistical inference. But its nature comes out most sharply with respect to testing hypotheses, so most of what will be said about it will be in that context.

The three types of questions

Let's consider the natures of conceptual, operational, and statistical questions.

The Scientific Question

A study for either scientific or decision-making purposes properly begins with a general question about the nature of the world—that is, a conceptual or theoretical question. One must then transform this question into an operational-empirical form that one can study scientifically. Thence comes the translation into a technical-statistical question.

The scientific-conceptual-theoretical question can be an issue of theory, or a policy choice, or the result of curiosity at large.

Examples include: Can a bioengineer increase the chance of female calves being born? Is copper becoming less scarce? Are the prices of liquor systematically different in states where the liquor stores are publicly owned compared to states where they are privately owned? Does a new formulation of pig rations lead to faster hog growth? Was the rate of unemployment higher last month than the long-run average, or was the higher figure likely to be the result of sampling error? What are the margins of probable error for an unemployment survey?

The Operational-Empirical Question

The operational-empirical question is framed in measurable quantities in a meaningful design. Examples include: How likely is this state of affairs (say, the new pig-food formulation) to cause an event such as was observed (say, the observed increase in hog growth)? How likely is it that the mean unemployment rate of a sample taken from the universe of interest (say, the labor force, with an unemployment rate of 10 percent) will be between 11 percent and 12 percent? What is the probability of

getting three girls in the first four children if the probability of a girl is .48? How unlikely is it to get nine females out of ten calves in an experiment on your farm? Did the price of copper fall between 1800 and the present? These questions are in the form of empirical questions, which have already been transformed by operationalizing from scientific-conceptual questions.

The Statistical Question

At this point one must decide whether the conceptual-scientific question is of the form of either a) or b):

a) A test about whether some sample will frequently happen by chance rather than being very surprising—a test of the "significance" of a hypothesis. Such hypothesis testing takes the following form: How likely is a given "universe" to produce *some* sample like x? This leads to interpretation about: How likely is a given universe to be the cause of *this observed* sample?

b) A question about the accuracy of the estimate of a parameter of the population based upon sample evidence (an inquiry about "confidence intervals"). This sort of question is considered by some (but not by me) to be a question in estimation—that is, one's best guess about (say) the magnitude and probable error of the mean or median of a population. This is the form of a question about confidence limits—how likely is the mean to be between x and y?

Notice that the statistical question is framed as a question in probability.

Illustrative translations

The best way to explain how to translate a scientific question into a statistical question is to illustrate the process.

Illustration A

Were doctors' beliefs as of 1964 about the harmfulness of cigarette smoking (and doctors' own smoking behavior) affected by the *social* groups among whom the doctors live (Simon, 1967-1968)? That was the *theoretical* question. We decided to define the doctors' *reference groups* as the *states* in which they live, because data about doctors and smoking were available state by state (*Modern Medicine*, 1964). We could then translate

this question into an operational and testable *scientific* hypothesis by asking this question: Do doctors in tobacco-economy states differ from doctors in other states in their smoking, and in their beliefs about smoking?

Which numbers would help us answer this question, and how do we interpret those numbers? We now were ready to ask the *statistical* question: Do doctors in tobacco-economy states "belong to the same universe" (with respect to smoking) as do other doctors? That is, do doctors in tobacco-economy states have the same characteristics—at least, those characteristics we are interested in, smoking in this case—as do other doctors? Later we shall see that the way to proceed is to consider the statistical hypothesis that these doctors do indeed belong to that same universe; that hypothesis and the universe will be called "benchmark hypothesis" and "benchmark universe" respectively—or in more conventional usage, the "null hypothesis."

If the tobacco-economy doctors do indeed belong to the benchmark universe—that is, if the benchmark hypothesis is correct—then there is a 49/50 chance that doctors in some state *other than* the state in which tobacco is most important will have the highest rate of cigarette smoking. But in fact we observe that the state in which tobacco accounts for the largest proportion of the state's income—North Carolina—had (as of 1964) a higher proportion of doctors who smoked than any other state. (Furthermore, a lower proportion of doctors in North Carolina than in any other state said that they *believed* that smoking is a health hazard.)

Of course, it is possible that it was just *chance* that North Carolina doctors smoked most, but the chance is only 1 in 50 if the benchmark hypothesis is correct. Obviously, *some* state had to have the highest rate, and the chance for any other state was also 1 in 50. But, because our original *scientific* hypothesis was that North Carolina doctors' smoking rate would be highest, and we then observed that it was highest even though the chance was only 1 in 50, the observation became interesting and meaningful to us. It means that the chances are strong that there was a connection between the importance of tobacco in the economy of a state and the rate of cigarette smoking among doctors living there (as of 1964).

To consider this problem from another direction, it would be rare for North Carolina to have the highest smoking rate for doctors if there were no special reason for it; in fact, it would occur only once in fifty times. But, if there *were* a special rea-

son—and we hypothesize that the tobacco economy provides the reason—then it would *not* seem unusual or rare for North Carolina to have the highest rate; therefore we choose to believe in the not-so-unusual phenomenon, that the tobacco economy caused doctors to smoke cigarettes.

Like many (most? all?) actual situations, the cigarettes and doctors' smoking issue is a rather messy business. Did I have a clear-cut, theoretically-derived prediction before I began? Maybe I did a bit of "data dredging"—that is, maybe I started with a vague expectation, and only arrived at my sharp hypothesis after I saw the data. This would weaken the probabilistic interpretation of the test of significance—but this is something that a scientific investigator does not like to do because it weakens his/her claim for attention and chance of publication. On the other hand, if one were a Bayesian, one could claim that one had a prior probability that the observed effect would occur, and the observed data strengthens that prior; but this procedure would not seem proper to many other investigators. The only wholly satisfactory conclusion is to obtain more data—but as of 1993, there does not seem to have been another data set collected since 1964, and collecting a set by myself is not feasible.

This clearly is a case of statistical inference that one could argue about, though perhaps it is true that *all* cases where the data are sufficiently ambiguous as to require a test of significance are also sufficiently ambiguous that they are properly subject to argument.

For some decades the hypothetico-deductive framework was the leading point of view in empirical science. It insisted that the empirical and statistical investigation should be preceded by theory, and only propositions suggested by the theory should be tested. Investigators were not supposed to go back and forth from data to theory to testing. It is now clear that this is an ivory-tower irrelevance, and no one lived by the hypothetico-deductive strictures anyway—just pretended to. Furthermore, there is no sound reason to feel constrained by it, though it strengthens your conclusions if you had theoretical reason in advance to expect the finding you obtained.

Illustration B

Does medicine CCC cure a cancer? That's the scientific question. So you give the medicine to six patients who have the cancer and you do not give it to six similar patients who have the cancer. Your sample contains only twelve people because

it is not feasible for you to obtain a larger sample. Five of six "medicine" patients get well, two of six "no medicine" patients get well. Does the medicine cure the cancer? That is, if future cancer patients take the medicine, will their rate of recovery be higher than if they did not take the medicine?

One way to translate the scientific question into a statistical question is to ask: Do the "medicine" patients *belong to the same universe* as the "no medicine" patients? That is, we ask whether "medicine" patients still have the *same* chances of getting well from the cancer as do the "no medicine" patients, or whether the medicine has bettered the chances of those who took it and thus removed them from the original universe, with its original chances of getting well. The original universe, to which the "no medicine" patients must still belong, is the benchmark universe. Shortly we shall see that we proceed by comparing the observed results against the benchmark *hypothesis* that the "medicine" patients still belong to the benchmark *universe*—that is, they still have the same chance of getting well as the "no medicine" patients.

We want to know whether or not the medicine does any good. This question is the same as asking whether patients who take medicine are still in the same population (universe) as "no medicine" patients, or whether they now belong to a different population in which patients have higher chances of getting well. To recapitulate our translations, we move from asking: Does the medicine cure the cancer? to, Do "medicine" patients have the same chance of getting well as "no medicine" patients?; and finally, to: Do "medicine" patients belong to the same universe (population) as "no medicine" patients? Remember that "population" in this sense does not refer to the population at large, but rather to a group of cancer sufferers (perhaps an infinitely large group) who have given chances of getting well, on the average. Groups with different chances of getting well are called "different populations" (universes). Shortly we shall see how to *answer* this statistical question. We must keep in mind that our ultimate concern in cases like this one is to *predict future results* of the medicine, that is, to predict whether use of the medicine will lead to a higher recovery rate than would be observed without the medicine.

Illustration C

Is method Alpha a better method of teaching reading than method Beta? That is, will method Alpha produce a higher

average reading score in the future than will method Beta? Twenty children taught to read with method Alpha have an average reading score of 79, whereas children taught with method Beta have an average score of 84. To translate this *scientific* question into a *statistical* question we ask: Do children taught with method Alpha come from the same universe (population) as children taught with method Beta? Again, "universe" (population) does *not* mean the town or social group the children come from, and indeed the experiment will make sense only if the children *do* come from the same population, in that sense of "population." What we want to know is whether or not the children belong to the same *statistical* population (universe), *defined according to their reading ability, after they have studied* with method Alpha or method Beta.

Illustration D

If one plot of ground is treated with fertilizer, and another similar plot is not treated, the benchmark (null) hypothesis is that the corn raised on the treated plot is no different than the corn raised on the untreated lot—that is, that the corn from the treated plot comes from ("belongs to") the same universe as the corn from the untreated plot. If our statistical test makes it seem very unlikely that a universe like that from which the untreated-plot corn comes would *also* produce corn such as came from the treated plot, then we are willing to believe that the fertilizer has an effect. For a psychological example, substitute the words "group of children" for "plot," "special training" for "fertilizer," and "I.Q. score" for "corn."

There is nothing sacred about the benchmark (null) hypothesis of "no difference." You could just as well test the benchmark hypothesis that the corn comes from a universe that averages 110 bushels per acre, if you have reason to be especially interested in knowing whether or not the fertilizer produces more than 110 bushels per acre. But in many cases it is reasonable to test the probability that a sample comes from the population that does not receive the special treatment of medicine, fertilizer, or training.

So far we have discussed the scientific question and the statistical question. Remember that there is always a generalization question, too: Do the statistical results from this particular sample of, say, rats apply to a universe of humans? This question can be answered only with wisdom, common sense, and general knowledge, and not with probability statistics.

Translating from a scientific question into a statistical question is mostly a matter of asking the probability that some given benchmark universe (population) will produce one or more observed samples. Notice that we must (at least for general scientific testing purposes) ask about a *given* universe whose composition we assume to be *known*, rather than about a *range* of universes, or about a universe whose properties are unknown. In fact, there is really only one question that probability statistics can answer: Given some particular benchmark universe of some stated composition, what is the probability that an observed sample would come from it? (Please notice the subtle but all-important difference between the words "would come" in the previous sentence, and the word "came.") A variation of this question is: Given two (or more) samples, what is the probability that they would come from the *same* universe—that is, that the same universe would produce both of them? In this latter case, the relevant benchmark universe is implicitly the universe whose composition is the two samples combined.

The necessity for stating the characteristics of the universe in question becomes obvious when you think about it for a moment. Probability-statistical testing adds up to comparing a sample with a particular benchmark universe, and asking whether there probably is a difference between the sample and the universe. To carry out this comparison, we ask *how likely* it is that the benchmark universe would produce a sample like the observed sample. But in order to find out whether or not a universe could produce a given sample, we must ask whether or not some *particular* universe—with stated characteristics—could produce the sample. There is no doubt that *some* universe could produce the sample by a random process; in fact, some universe did. The only sensible question, then, is whether or not a *particular* universe, with stated (or known) characteristics, is likely to produce such a sample. In the case of the medicine, the universe with which we compare the sample who took the medicine is the benchmark universe to which that sample would belong if the medicine had had no effect. This comparison leads to the benchmark (null) hypothesis that the sample comes from a population in which the medicine (or other experimental treatment) seems to have *no effect*. It is to avoid confusion inherent in the term "null hypothesis" that I replace it with the term "benchmark hypothesis."

The concept of the benchmark (null) hypothesis is not easy to grasp. The best way to learn its meaning is to see how it is used in practice. For example, we say we are willing to be-

lieve that the medicine has an effect if it seems very unlikely from the number who get well that the patients given the medicine still belong to the same benchmark universe as the patients given no medicine at all—that is, if the benchmark hypothesis is unlikely.

The steps in statistical inference

These are the steps in conducting statistical inference

Step 1. Frame a question in the form of: What is the chance of getting the observed sample x from some specified population X? For example, what is the probability of getting a sample of 9 females and one male from a population where the probability of getting a single female is .48?

Step 2. Reframe the question in the form of: What kinds of samples does population X produce, with which probabilities? That is, what is the probability of the observed sample x (9 females in 10 calves), given that a population is X (composed of 48 percent females)? Or in notation, what is $p(x \mid X)$?

Step 3. Actually investigate the behavior of S with respect to S and other samples. This can be done in two ways:

a. Use the calculus of probability (the formulaic method), perhaps resorting to the Monte Carlo method if an appropriate formula does not exist. Or

b. Resampling (in the larger sense), which equals the Monte Carlo method minus its use for approximations, investigation of complex functions in statistics and other theoretical mathematics, and non-resampling uses elsewhere in science. Resampling in the more restricted sense includes bootstrap, permutation, and other non-parametric methods. More about the resampling procedure follows in the paragraphs to come, and then in later chapters in the book.

Step 4. Interpret the probabilities that result from step 3 in terms of acceptance or rejection of hypotheses, surety of conclusions, and as inputs to decision theory.

The following short definition of statistical inference summarizes the previous four steps: Statistical inference equals the selection of a probabilistic model to resemble the process you

wish to investigate, the investigation of that model's behavior, and the interpretation of the results.

Stating the steps to be followed in a procedure is an operational definition of the procedure. My belief in the clarifying power of this device (the operational definition) is embodied in the set of steps given in Chapter 10 for the various aspects of statistical inference. A canonical question-and-answer procedure for testing hypotheses will be found in Chapter 19, and one for confidence intervals will be found in Chapter 20.

Summary

We define resampling to include problems in inferential statistics as well as problems in probability as follows: *Using the entire set of data you have in hand, or using the given data-generating mechanism (such as a die) that is a model of the process you wish to understand, produce new samples of simulated data, and examine the results of those samples*. That's it in a nutshell. In some cases, it may also be appropriate to amplify this procedure with additional assumptions.

Problems in pure probability may at first seem different in nature than problems in statistical inference. But the same logic as stated in this definition applies to both varieties of problems. The difference is that in probability problems the "model" is known in advance—say, the model implicit in a deck of poker cards plus a game's rules for dealing and counting the results—rather than the model being assumed to be best estimated by the observed data, as in resampling statistics.

The hardest job in using probability statistics, and the most important, is to *translate* the scientific question into a form to which statistics can give a sensible answer. You must translate scientific questions into the appropriate form for *statistical operations*, so that you know which operations to perform. This is the part of the job that requires hard, clear thinking—though it is non-mathematical thinking—and it is the part that someone else usually cannot easily do for you.

Once you know exactly which probability-statistical question you want to ask—that is, exactly which probability you want to determine—the rest of the work is relatively easy. The stage at which you are most likely to make mistakes is in stating the

question you want to answer in probabilistic terms. Though this step is hard, *it involves no mathematics*. This step requires only *hard, clear thinking*. You cannot beg off by saying "I have no brain for math!" To flub this step is to admit that you have no brain for clear thinking, rather than no brain for mathematics.

Endnotes

1. These steps are discussed in more philosophic depth in my forthcoming book on the philosophy of statistics and resampling.