Chapter 5. EXPERIMENTS, SAMPLING, AND STATISTICAL DECISIONS

1. Experiments

The business of economics is to explain how consumers and firms behave, and the implications of this behavior for the operation of the economy. To do this, economists need to be able to describe the features of the economy and its economic agents, to model behavior, and to test the validity of these models. For example, economists are interested in determining the rate of employment of unwed mothers, and how this is affected by changes in government welfare programs for low-income mothers with dependent children. In principle, one could take a census of the populational to determine this employment rate.

In the example of unwed mothers, we believe that economic incentives implicit in welfare programs influence willingness to work, so that changes in these programs may *cause* the employment rate to change. Our investigation may seek to determine this causal effect, rather than simply providing a snapshot of the state of employment in this population. While obtaining accurate descriptions of population features is important in economics, most econometric analysis is devoted to measuring causal effects.

The most reliable way to study a causal relationship is through an *experiment*. For example, to study the effect of welfare programs on employment, one could in principle establish several different levels of welfare support, or *treatments*, and assign these treatments at random to members of the population. The measured response of employment to these treatment is the causal effect we were looking for, with the random assignment of treatments assuring that the effects we see are arising from this source alone, not from other, uncontrolled factors that might happen to be correlated with the treatment. A census before and after the change would reveal the impact of the treatments on the employment rate, provided there are no other uncontrolled elements that could confound the comparison.

Economists only rarely have the freedom to study economic relationships by conducting classical experiments with design and random assignment of treatments. At the scale of economic policy, it would often be invasive, time-consuming, and costly to conduct the experiments one would like. In addition, being experimented on can make economic agents and economies testy. This makes economics primarily an observational or *field* science, like astronomy, rather than a *bench* science, like chemistry. Economists must search for *natural experiments* in which economic agents are subjected to varying levels of the causal factor under circumstances where the effects of potential confounding factors are controlled, either by something like random assignment of treatments or by measuring the levels of potential confounding factors and using modeling and data analysis methods that can untangle the separate effects of different factors. For example, to study the impact of welfare programs on the employment of unwed mothers, we might treat the adoption of different welfare laws by different States at different times as a natural experiment. This is not as clean as random assignment of treatments, because economic circumstances differ across States and this may influence what welfare programs are adopted. Then, one is left with the problem of determining how much of a change in employment rate between States with strong and weak work incentives in their welfare programs is due to these incentives, and how much is due to overall demographics or income levels that induced States to adopt one welfare program or the other.

Looking for good natural experiments is an important part of econometric analysis. The most persuasive econometric studies are those where Nature has provided an experiment in which there is little possibility than anything other than the effect you are interested in could be causing the observed response. In data where many factors are at work jointly, the ability of statistical analysis to identify the separate contribution of each factor is limited. Regression analysis, which forms the core of econometric technique, is a powerful tool for partialling out the contributions of different factors, but even so it is rarely definitive. A good way to do econometrics is to look for good natural experiments *and* use statistical methods that can tidy up the confounding factors that Nature has not controlled for us.

2. Populations and Samples

Often, a population census is impractical, but it is possible to *sample* from the population. A core idea of statistics is that a properly drawn sample is a representation of the population, and that one can exploit the analogies between the population and the sample to draw inferences from the sample about features of the population. Thus, one can measure the rate of employment of unwed mothers in the sample, and use it to infer the rate of employment of unwed mothers in the population. Statistics provides the tools necessary to develop these analogies, and assess how reliable they are.

A basic statistical concept is of a *simple random sample*. The properties of a simple random sample are that every member of the population has the same probability of being included, and the sample observations are statistically independent. A simple random sample can be defined formally in terms of independent trials from a probability space; see Chap. 3.4. However, for current purposes, it is sufficient to think of a population that is characterized by a probability distribution, and think of a random sample as a sequence of observations drawn independently from this distribution.

A simple random sample is representative of the underlying population in the sense that each sample observation has the population probability distribution. However, there is a more fundamental sense in which a simple random sample is an analog of the population, so that sample statistics are appealing approximations to their population analogs. Suppose one is dealing with a one-dimensional random variable X that is distributed in the population with a CDF F(x), and that one is interested in some feature of this distribution, such as its mean $\mu = \int_{-\infty}^{+\infty} x\cdot dF$.[1] This expectation depends on F, and we could make the dependence explicit by writing it as $\mu(x,F)$. More generally, the target may be $\mu(g,F)$, where $g = g(x)$ is some function of x, such as $g(x) = x^2$ or $g(x) = \mathbf{1}(x\le 0)$.

Now suppose $(x_1,...,x_n)$ is a simple random sample drawn from this CDF, and define

$$F_n(x) = \sum_{i=1}^{n} \mathbf{1}(x_i \le x) \ .$$

Then $F_n(x)$ equals the fraction of the sample values that are no greater than x. This is called the *empirical CDF* of the sample. It can be interpreted as coming from a probability measure that puts weight 1/n on each sample point. A sample analog of the population mean μ is the sample mean, which is usually written as $\frac{1}{n} \sum_{i=1}^{n} x_i$, but can also

[1] Recall from Chap. 3.5 that the notation for an expectation of a function g(X) with respect to a probability measure ν is $\mathbf{E}g(X) = \int g(x)\cdot\nu(dx)$. When the probability measure is represented by a CDF F(x), we denote this expectation by $\int g(x)\cdot F(dx)$ or by $\int g(x)\cdot dF$. This notation encompasses both CDF's that have density functions with respect to Lebesgue measure, where the integral becomes $\int g(x)\cdot F'(x)dx$, and CDF's that have finite or countable support, as well as mixed cases. In the example of employment of unwed mothers, the variable x is discrete, with $x = 1$ if employed and $x = 0$ otherwise. In this example, $\mathbf{E}g(X) = g(0)\cdot F(0) + g(1)\cdot(1 - F(0))$.

be written as $\mu(x,F_n) = \int_{-\infty}^{+\infty} x\cdot dF_n$. This notation emphasizes that the sample mean is a function of the empirical CDF of the sample. The only difference in the population mean and the sample mean is that the first is a function of F and the second is the same function of F_n.

The following proposition, sometimes called the "fundamental theorem of statistics", establishes that as a simple random sample gets larger and larger, its empirical CDF approximates the population CDF more and more closely:

Glivenko-Cantelli Theorem. *If random variables* $X_1,X_2,...$ *are independent and have a common CDF* F, *then* $\sup_x|F_n(x) - F(x)|$ *converges to zero almost surely as* $n \to \infty$.

Proof: Given ε, $\delta > 0$, there exists a finite number of points $z_1 < ... < z_K$ such that the monotone right-continuous function F varies at most ε between the points; i.e., $F(z_k^-) - F(z_{k-1}) < \varepsilon/2$.[2] For every x, bracketed by $z_{k-1} \le x < z_k$, one has

$$F_n(z_{k-1}) - F(z_{k-1}) - \varepsilon/2 \le F_n(x) - F(x) \le F_n(z_k) - F(z_k) + \varepsilon/2.$$

The event $\sup_k |F_n(z_k) - F(z_k)| < \varepsilon/2$ implies the event $\sup_x |F_n(x) - F(x)| < \varepsilon$. At each z_k, the Kolmogorov Strong Law of Large Numbers establishes that $F_n(z_k) \xrightarrow{as} F(z_k)$; i.e., there exists n_k such that $P(\sup_{n\ge n_k}|F_n(z_k) - F(z_k)| > \varepsilon/3) < \delta/K$. Let $n_o = \max_k n_k$. Then with probability at least $1 - \delta$, the event $\sup_{n\ge n_o} \sup_k|F_n(z_k) - F(z_k)| < \varepsilon/2$, implying the event $\sup_{n\ge n_o} \sup_k |F_n(z_k) - F(z_k)| < \varepsilon$, occurs. □

[2] Any point where F jumps by more than $\varepsilon/4$ will be included as a z_k point. By convention, assume $z_1 = -\infty$ and $z_K = +\infty$.

An implication of the Glivenko-Cantelli theorem is that if $\mu(g,\cdot)$ is continuous at F, then the sample statistic $\mu(g,F_n)$ converges almost surely to the population statistic $\mu(g,F)$.[3] This provides a fundamental justification for the use of simple random samples, and for the use of sample statistics $\mu(g,F_n)$ that are analogs of population statistics $\mu(g,F)$ that are of interest.

While the idea of simple random sampling is straightforward, implementation in applications may not be. The way sampling is done is to first establish a *sample frame* and a *sampling protocol*. The sample frame essentially identifies the members of the population in an operational way that makes it possible for them to be sampled. For example, suppose your target population is the population of unwed mothers in the U.S. A sample frame could be a master list containing the names and telephone numbers of all unwed mothers. The sampling protocol in this case could be to call individuals from this list with equal probability, using a random number generator. However, this is impractical because the master list does not exist. A practical sample frame would be the list of all working residential telephone numbers in the U.S. The sampling protocol would be to call numbers from this list with equal probability, ask screening questions to determine if any unwed mothers live at that number, and interview an eligible resident if there is one. This would yield a sample that is not exactly a simple random sample, because some unwed mothers may not have telephones, households

[3] We have not at this point defined precisely what it means for a function $\mu(g,\cdot)$ to be continuous on a space of functions that includes the population and empirical CDF's. For current discussion, it is enough to note that if g is a continuous function of x and $\lim_{M\to\infty} \int_{|x|>M} |g(x)|\cdot dF \longrightarrow 0$ so that $\mathbf{E}g(X)$ exists, then this continuity requirement is met and $\mu(g,F_n) \xrightarrow{as} \mu(g,F)$.

with multiple telephones are oversampled relative to those with one telephone, some households may contain more than one unwed mother, and there may be attrition because some telephones are not answered or the respondent declines to participate. An important aspect of econometric analysis is determining when deviations from simple random sampling matter, and developing methods for dealing with it.

There are a variety of sampling schemes that are more complex variants on simple random sampling, with protocols that produce various forms of stratification. An example is cluster sampling, which first samples geographical units (e.g., cities, census tracts), and then samples residences within each chosen unit. Generally, these schemes are used to reduce the costs of sampling. Samples produced by such protocols often come with sample weights, the idea being that when these are applied to the sample observations, sample averages will be reasonable approximations to population averages. Under some conditions, econometric analysis can be carried out on these stratified samples by treating them *as if* they were simple random samples. However, in general it is important to consider the implications of sampling frames and protocols when one is setting up a statistical analysis.

We have given a strong theoretical argument that statistical analysis of simple random samples will give reasonable approximations to target population features. On the other hand, the history of statistics is filled with horror stories where an analysis has gone wrong because the sample was not random. The classical example was the Liberty telephone poll in 1936 that predicted that Roosevelt would lose the election. Roosevelt won in a landslide. The problem was that only the rich had telephones. One should be very skeptical of statistical analyses that use purposive or selected samples, as the safeguards provided by random sampling no longer apply and sample statistics may be poor approximations to population statistics.

An arena where sampling is particularly obscure is in analysis of economic time-series. Here, one is observing a slice of history, and the question is what is the population and in what sense is this slice a suitable sample? One way statisticians have thought about this is to visualize "parallel universes", with our universe being one realization. This works for doing the formal probability theory, but is unsatisfying for the economist whose target is a hypothesis about this universe, not about the population of universes. Another way to approach the problem is to think about the time series sample as a slice of a stochastic process that operates through time, with certain rules that regulate the relationship between behavior in a slice and behavior through all time. For example, one might postulate that the stochastic process is *stationary* and *ergodic*, which would mean that the distributions of variables depend only on their relative position in time, not their absolute position, and that long run averages converge to limits.

In this chapter and several chapters following, we will assume that the samples we are dealing with are simple random samples. Once we have a structure for statistical inference in this simplest case, we will turn to the problems that arise under alternative sampling protocols.

3. Statistical Decisions

The process of statistical estimation can be thought of as decision-making under uncertainty. The economic problem faced by Petit Verdot in Chapter 1 is an example. In decision-making under uncertainty, one has limited information, based upon observed data. There are costs to mistakes. On the basis of the available information, one wants to choose an action that minimizes cost. Let **x** denote the data, which may be a vector of observations from a simple random sample, or some more complex sample such as a slice from a time series process. These observations are governed by a *probability law*, or *data generation process* (DGP). We do not know the true DGP, but assume now that we do know that it is a member of some family of possible DGP's which we will index by a parameter θ. The true DGP will correspond to a value θ_o of this index. Let $F(\mathbf{x},\theta)$ denote the CDF for the DGP corresponding to the index θ. For the remainder of this discussion, we will assume that this CDF has a density, denoted by $f(\mathbf{x},\theta)$.[4] The density f is called the *likelihood* function of the data. The unknown parameter θ_o might be some population feature, such as the frequency of employment in the unwed mother example. The statistical decision problem might then be to estimate θ_o, taking into account the cost of errors. Alternately, θ_o might be one of two possible values, say 0 and 1, corresponding to the DGP an economist would expect to see when a particular hypothesis is true or false, respectively. In this case, the decision problem is to infer whether the hypothesis is in fact true, and again there is a cost of making an error.

Where do we get values for the costs of mistakes? If the client for the statistical analysis is a business person or a policy-maker, an inference about θ_o might be an input into an action that has a payoff in profits or in a measure of social welfare that is indexed in dollars. A mistake will lower this payoff. The cost of a mistake is then the opportunity cost of foregoing the higher payoff to be obtained if one could make an optimal decision. For the example of employment of unwed mothers, making a mistake on the employment rate may cause the planned welfare budget to go out of balance, and cost may be a known function of the magnitude of the unanticipated

[4] We could associate a family of DGP's with a family of probability measures on the space containing **x**, but we will not need this level of generality. Chapter 3.5.5 establishes that if a probability measure is absolutely continuous with respect to a length measure, such as Lebesgue or counting measure, on the space containing **x**, then a density defined with respect to the length measure exists.

imbalance. However, if there are multiple clients, or the analysis is being performed for the scientific community, there may not be precise costs, and it may be necessary to provide sufficient information from the analysis so that most potential users can determine their most appropriate action. Before considering this situation, we will look at the case where there is a known cost function $C(\theta,\theta_o)$ that depends on the true parameter value θ_o and on the inference θ made from the data.

A decision rule, or *action*, will be a mapping $T(\cdot)$ from the data **x** into the space of possible θ values. Note that while $T(\mathbf{x})$ depends on the data **x**, it can depend on the unknown parameter θ_o only through the influence of θ_o on the determination of **x**, not directly. Because the data are random variables, $T(\cdot)$ is also a random variable, and it will have a density $\psi(t,\theta_o)$ given by a transformation of $f(\mathbf{x},\theta_o)$. The cost associated with the action $T(\cdot)$, given data **x**, is $C(T(\mathbf{x}),\theta_o)$. One would like to choose this to be as small as possible, but the problem is that one cannot do this without knowing θ_o. However, the client may, prior to the observation of **x**, have some beliefs about the likely values of θ_o. We will assume that these *prior beliefs* can be summarized in a density $h(\theta)$. Given this prior belief, it is possible to calculate an expected cost for an action $T(\cdot)$. First, apply Bayes law to the *joint* density $f(x,\theta_o)\cdot h(\theta_o)$ of **x** and θ_o to obtain the conditional density of θ_o given x,

$$(1) \qquad p(\theta_o|\mathbf{x}) = f(\mathbf{x},\theta_o)\cdot h(\theta_o)/\int_{-\infty}^{+\infty} f(\mathbf{x},\theta)\cdot h(\theta)d\theta \ .$$

This is called the *posterior* density of θ_o, given the data **x**. Using this posterior density, one can calculate the expected cost for an action $T(\mathbf{x})$:

$$(2)\qquad \mathbf{E}_{\theta_o|\mathbf{x}}\, C(T(\mathbf{x}),\theta_o) = \int_{-\infty}^{+\infty} C(T(\mathbf{x}),\theta_o)\cdot f(\mathbf{x},\theta_o)\cdot h(\theta_o)d\theta_o / \int_{-\infty}^{+\infty} f(\mathbf{x},\theta)\cdot h(\theta)d\theta \ .$$

The expected cost in (2) is called the *Bayesian risk*. It depends on the function T(·). The optimal action $T^*(\cdot)$ is the function T(·) that minimizes the expected cost for each **x**, and therefore minimizes the Bayesian risk.

In general, it is not obvious what the optimal action $T^*(\mathbf{x})$ that minimizes (2) looks like. A few examples help to provide some intuition:

(i) Suppose $C(\theta,\theta_o) = (\theta - \theta_o)^2$, a quadratic cost function in which cost is proportional to the square of the distance of the estimator T(**x**) from the true value θ_o. For a given **x**, the value θ = T(**x**) that minimizes (2) has to satisfy the first-order condition $0 = \int_{-\infty}^{+\infty} (T^*(\mathbf{x}) - \theta_o)\cdot f(\mathbf{x},\theta_o)\cdot h(\theta_o)d\theta_o$, or

$$(3)\qquad T^*(\mathbf{x}) = \int_{-\infty}^{+\infty} \theta_o\cdot f(\mathbf{x},\theta_o)\cdot h(\theta_o)d\theta_o / \int_{-\infty}^{+\infty} f(\mathbf{x},\theta)\cdot h(\theta)d\theta = \int_{-\infty}^{+\infty} \theta\cdot p(\theta|\mathbf{x})\cdot d\theta \ .$$

Then, $T^*(\mathbf{x})$ equals the *mean of the posterior density*.

(ii) Suppose $C(\theta,\theta_o) = \alpha\cdot\max(0,\theta - \theta_o) + (1-\alpha)\cdot\max(0,\theta_o - \theta)$ where α is a cost parameter satisfying 0 < α < 1. This cost function is linear in the magnitude of the error. When α = 1/2, the cost function is symmetric; for smaller α it is unsymmetric with a unit of positive error costing less than a unit of negative error. The first-order condition for minimizing cost is

$$0 = -(1-\alpha)\cdot\int_{-\infty}^{T^*(\mathbf{x})} f(\mathbf{x},\theta_o)\cdot h(\theta_o)d\theta_o + \alpha\cdot\int_{T^*(\mathbf{x})}^{+\infty} f(\mathbf{x},\theta_o)\cdot h(\theta_o)d\theta_o ,$$

or letting $P(\theta|\mathbf{x})$ denote the CDF of the posterior density, $P(T^*(\mathbf{x})|\mathbf{x}) = \alpha$. Then $T^*(\mathbf{x})$ equals the α*-level quantile* of the posterior distribution. In the case that $\alpha = 1/2$, so that costs are symmetric in positive and negative errors, this criterion picks out the *median of the posterior density*.

(iii) Suppose $C(\theta,\theta_o) = -1/2\alpha$ for $|\theta-\theta_o| \le \alpha$, and $C(\theta,\theta_o) = 0$ otherwise. This is a cost function that gives a profit of $1/2\alpha$ when the action is within a distance α of θ_o, and zero otherwise; α is a positive cost parameter. The criterion (2) requires that $\theta = T^*(\mathbf{x})$ be chosen to minimize the expression $-\frac{1}{2\alpha}\cdot\int_{\theta-\alpha}^{\theta+\alpha} p(\theta_o|\mathbf{x})\cdot d\theta_o$. Now suppose α is very small. Then the criterion equals $-p(\theta|\mathbf{x})$ to a very close approximation. Then the right-hand-side of this expression is approximately $-p(\theta|\mathbf{x})$. The argument minimizing $-p(\theta|\mathbf{x})$ is called the *maximum posterior likelihood* estimator; it picks out the *mode of the posterior density*. Then for α very small, the optimal estimator is approximately the maximum posterior likelihood estimator. Recall that $p(\theta|\mathbf{x})$ is proportional to $f(\mathbf{x},\theta)\cdot h(\theta)$. Then, the first-order condition for its maximization can be written

(4) $$0 = \frac{\nabla_\theta f(\mathbf{x},\theta)}{f(\mathbf{x},\theta)} + \frac{\nabla_\theta h(\theta)}{h(\theta)} .$$

The first term on the right-hand-side of this condition is the derivative of the log of the likelihood function, also called the *score*. The second term is the derivative of the log of the prior density. If prior beliefs are strong and tightly concentrated, then the second term will be very important, and the maximum will be close to the mode of the prior density, irrespective of the data. On the other hand, if prior beliefs are weak and very disperse, the second term will be small and the maximum will be close to the mode of the likelihood function. In this limiting case, the solution to the statistical decision problem will be close to a general-purpose classical estimator, the maximum likelihood estimator.

The cost function examples above were analyzed under the assumption that prior beliefs were characterized by a density with respect to Lebesgue measure. If, alternately, the prior density had a finite support, then one would have analogous criteria, with sums replacing integrals, and the criteria would pick out the best point from the support of the prior.

The idea that there are prior beliefs regarding the true value of θ_o, and that these beliefs can be characterized in terms of a probability density, is called the *Bayesian* approach to statistical inference. It is philosophically quite different than an approach that thinks of probabilities as being associated with particular random devices such as coin tosses that can produce frequencies. Bayesian statistics assumes that humans have a coherent system of beliefs that can attach probabilities to events such as "the Universe will continue to expand forever" and "the employment rate for unwed mothers is no higher than 62 percent", and these personal probabilities satisfy the basic axioms of probability theory. One of the implications of this way of thinking is that it is meaningful to talk about the *probability* that something (an event) is true. This will be important in how one thinks about an economic hypothesis, such as the hypothesis that the employment rate of unwed mothers does not depend on the wage tax implicit in a welfare policy. In classical statistics, this hypothesis is either true or false, and the purpose of statistical inference is to decide whether it is true. In Bayesian statistics, this would correspond to concluding that the probability that the event is true is either zero or one. For a Bayesian statistician, it is rarely possible to be this certain, and it is much more meaningful to talk about the probability of the the event being high or low.

The statistical decision theory just developed assumed that the analysis had a client with precisely defined prior beliefs. As in the case of the cost of errors, there will be circumstances where the client's prior beliefs are not known, or there is not even a well-defined client. It is the lack of a clearly identified prior that is one of the primary barriers to acceptance of the Bayesian approach to statistics.[5] There are three possible options in the situation where there is not a well-defined prior. One is to stop the analysis short of a final solution, and simply deliver sufficient information from the sample to enable each potential user to compute the action appropriate to her own cost function and prior beliefs. This can be accomplished by delivering a set of *sufficient statistics* for the problem. We will later give a formal definition of sufficient statistics, but informally they are any reduction of the data **x** that still contains all the information in **x** that is useful in drawing inferences on θ. The limitations of this approach are that the dimensionality of sufficient statistics can be high, in many cases the dimension of the full sample, and that a substantial computational burden is being imposed on the user.

The second option is to carry out the statistical decision analysis with prior beliefs that carry "zero information". For example, an analysis may use a "diffuse" prior that gives every value of θ an equal probability. There are some technical problems with this approach. If the set of possible θ values is unbounded, "diffuse"

[5] Bayesain computations can be quite difficult, and this is a second major barrier.

priors may not be proper probability densities that integrate to one. This problem can be overcome by using the prior without normalization, or by forming it as a limit of proper priors. More seriously, the idea of equal probability as being equivalent to "zero information" is flawed. A one-to-one but nonlinear transformation of the index θ can change a "diffuse" prior with equal probabilities into a prior in which probabilities are not equal, without changing anything about available information or beliefs. Then, equal probability is not in fact a characterization of "zero information". The technique of using diffuse or "uninformed" priors is fairly popular, in part because it simplifies some calculations. However, one should be careful to not assume that an analysis based on a particular set of diffuse priors is "neutral" or "value-free".

The third option is based on the idea that you are in a game against Nature in which Nature plays θ_o and reveals the information **x** about her strategy that you know is a draw from the DGP $f(\mathbf{x},\theta_o)$, and you play $T(\mathbf{x})$. The expectation $R(T(\cdot),\theta_o) \equiv \mathbf{E}_{\mathbf{x}|\theta_o}C(T(\mathbf{x}),\theta_o) = \int_{-\infty}^{+\infty} C(T(\mathbf{x}),\theta_o)\cdot f(\mathbf{x},\theta_o)d\mathbf{x}$ is called the *risk function*. A strategy $T'(\mathbf{x})$ is called *inadmissible* if there is a second strategy $T''(\mathbf{x})$ such that if $R(T''(\cdot),\theta_o) \leq R(T'(\cdot),\theta_o)$ for all θ_o, with the inequality strict for some θ_o. It is always in your interest to avoid inadmissible strategies. This falls short of a full precription for decision-making because different strategies need not be comparable in terms of this dominance criterion, and there are typically many admissible strategies. A conservative strategy in games is to play in such a way that you minimize the maximum cost your opponent can impose on you. Taking the risk function as the measure of cost, this strategy chooses $T(\cdot)$ to minimize $\max_{\theta_o}R(T(\cdot),\theta_o)$. This is called a *minimax* strategy. It guarantees that your risk will be no higher than $\min_{T(\cdot)} \max_{\theta_o}R(T(\cdot),\theta_o)$.

This is a sensible strategy in a zero-sum game with a clever opponent, since your cost is your opponent's gain. It is not obvious that it is a good strategy in a game against Nature, since the game is not necessarily zero-sum and it is unlikely that Nature is an aware opponent who cares about your costs.

4. Statistical Inference

Statistical decision theory provides a template for statistical analysis when it makes sense to specify prior beliefs and costs of mistakes. Its emphasis on using economic payoffs as the criterion for statistical inference is appealing to economists as a model of decision-making under uncertainty, and provides a comprehensive, but not necessarily simple, program for statistical computations. While the discussion in this chapter concentrated on estimation questions, we shall see in Chapter 7 that it is also useful for considering tests of hypotheses.

The primary limitation of the Bayesian analysis that flows from statistical decision theory is that it is difficult to rationalize and implement when costs of mistakes or prior beliefs are not fully spelled out. In particular, in most scientific work where the eventual user of the analysis is not identified, so there is no concensus on costs of mistakes or priors, there is often a preference for "purely objective" solutions rather than Bayesian ones. Since a Bayesian approach can in principle be structured so that it provides solutions for all possible costs and priors, including those of any prospective user, this preference may seem puzzling. However, there may be compelling computational reasons to turn to "classical" approaches to estimation as as alternative to the Bayesian statistical decision-making framework. We will do this in the next chapter.