CHAPTER 5. EXPERIMENTS, SAMPLING, STATISTICAL DECISIONS

5.1. 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 effects of medicare eligibility on retirement decisions. They believe that the incentives implicit in medical insurance programs influence willingness to work, so that changes in these programs may *cause* retirement behavior to change. A first level of empirical interest is a description of the current situation, a snapshot of current retirement patterns under the current program. This description could be based on a census or sample of the current population. Statistics will play a role if a sample is used, providing tools for judging the accuracy of estimates of population parameters. At a deeper level, economists want to estimate how patterns would change if eligibility rules were altered. This interest requires that one conduct, or at least observe, an "experiment" in which different workers face different programs, and the impact of the program differences on their responses can be observed. The objective is to uncover a *causal* mapping from programs to behavioral responses. Major barriers to accomplishing this are confounding causal factors, or mutual causation by deeper hidden effects. A well-designed experiment or intelligent use of a "natural experiment" that provides a clear separation between the factor of interest and possible confounding effects will provide the most compelling empirical evidence on the economic question.

The most reliable way to try to uncover a causal relationship is through a *designed experiment*. For example, to study the effect of the medicare program on retirement, one could in principle establish several different levels of medicare eligibility, 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. Classical prototypes for designed experiments are those done in chemistry or biology labs, where a good procedure will be effective in eliminating potential confounding factors so the effect of the one factor of interest can be measured. Even here, there can be problems of measurement error and contaminated experiments, and statistical issues arise. Perhaps better prototypes for experiments in economics are designed field experiments in ecology or agronomy. For example, consider the classical experiment to measure the impact of fertilization on the productivity of corn plants. The agronomist prepares different test plots, and tries to keep conditions other than fertilizer, such as irrigation levels, comparable across the plots. However, there will be a variety of factors, such as wind and sunshine, that may differ from one plot to another. To isolate the effect of fertilizer from these confounding effects, the agronomist assigns the fertilizer treatments to the different plots at random. This *randomized treatment design* is a powerful tool for measuring the causal effect of the treatments. Economists rarely have the freedom to study economic relationships by designing classical experiments with 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 need. In addition, being experimented on can make economic agents and economies testy. However, there are various arenas in which designed experiments are done in economics. Field experiments have examined the impact of different marginal tax rates on employment behavior in low-income families, and the effect of different job training programs. Economics laboratory experiments have studied behavior in artificial markets. However, some areas of economic interest are beyond the reach of designed experiments because of technical and ethical barriers. No one would seriously propose, for example, to study the effect of life expectancy on savings behavior by randomly assigning execution dates, or the returns to education by randomly assigning years of schooling that students may receive. This makes economics primarily an observational or *field* science, like astronomy. Economists must search for *natural experiments* in which economic agents are subjected to varying levels of a causal factor of interest under circumstances where the effects of potential confounding factors are controlled, either by something like random assignment of treatments by Nature, 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 schooling on income, we might try to use as a "Natural experiment" individuals such as Vietnam war draftees and non-draftees who as a result of the random draft lottery have different access to schooling. To study the impact of medical insurance eligibility on retirement decisions, we might try to study individuals in different states where laws for State medical welfare programs differ. 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 variation in retirement patterns between States with strong and weak work incentives in their medical 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 separating 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.

5.2. POPULATIONS AND SAMPLES

5.2.1. 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 average retirement age in the sample, and use it to infer the mean retirement age 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.

5.2.2. 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 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 $\mu = \int_{-\infty}^{+\infty} x\cdot F(dx)$. 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

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

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. The population mean $\mu(x,F)$ has a sample analog which is usually written as $\bar{x} = n^{-1}\sum_{i=1}^{n} x_i$, but can also be written as $\mu(x,F_n) = \int_{-\infty}^{+\infty} x\cdot F_n(dx)$. This notation emphasizes that the sample mean is a function of the empirical CDF of the sample. The population mean and the sample mean are then the same function $\mu(x,\cdot)$, the only difference being that the first is evaluated at F and the second is evaluated at 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. Then, intuitively, if $\mu(g,\cdot)$ is "continuous" in its second argument, an analogy principle suggests that $\mu(g,F_n)$ will converge to $\mu(g,F)$, so that $\mu(g,F_n)$ will be a good estimator of $\mu(g,F)$.

Theorem 5.1. (Glivenko-Cantelli) 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.

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 $\varepsilon/2$ between the points; i.e., $F(z_k{*}) - F(z_{k-1}) < \varepsilon/2$, where $z_k{*}$ denotes the limiting value as one approaches z_k from the left. 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 $z_K = +\infty$. 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 $\sup_k|F_n(z_k) - F(z_k)| < \varepsilon/2$ then implies the event $\sup_x |F_n(x) - F(x)| < \varepsilon$. At each z_k, the Kolmogorov SLLN establishes that $F_n(z_k) \to_{as} F(z_k)$. Then there exists n_k such that the probability of $|F_n(z_k) - F(z_k)| > \varepsilon/3$ for any $n \ge n_k$ is less than δ/K. Let $n = \max_k n_k$. Then, with probability at least $1 - \delta$, the event $\sup_{m>n} \sup_k |F_m(z_k) - F(z_k)| < \varepsilon/2$ occurs, implying the event $\sup_{m>n} \sup_x |F_m(x) - F(x)| < \varepsilon$, occurs. □

The Glivenko-Cantelli theorem implies that F_n converges in distribution to F, but is stronger, establishing that the convergence is uniform rather than pointwise, and is not restricted to continuity points of F. It is useful to state the Kolmogorov SLLN in the terminology used here: If the population statistic $\mu(g,F)$ exists, then the sample statistic $\mu(g,F_n)$ converges almost surely to $\mu(g,F)$. 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.

5.2.3. 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, and the sampling protocol spells out precisely how the sampling is to be done and the data collected. For example, suppose your target is the population of individuals who were 55 years of age in 1980 and over the following twenty years have retired in some pattern that may have been influenced by their access to medical insurance. An ideal sample frame would be a master list containing the names and current telephone numbers of all individuals in this target population. The sampling protocol could then be to use a random number generator to select and call individuals from this list with equal probability, and collect data on their retirement age and economic circumstances. However, the required master list does not exist, so this simple sample design is infeasible. A practical sample frame might instead start from a list of all working residential telephone numbers in the U.S. The sampling protocol would be to call numbers at random from this list, ask screening questions to determine if anyone from the target population lives 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 members of the target population have died or do not have telephones, households with multiple telephones are over sampled relative to those with one telephone, some households may contain more than one eligible person, and there may be attrition because some telephones are not answered or the respondent declines to participate. Even this sampling plan is infeasible if there is no master list of all the working residential telephone numbers. Then one might turn instead to random digit dialing (RDD), with a random number generator on a computer making up potential telephone numbers at random until the phone is answered. At first glance, it may seem that this is guaranteed to produce at least a simple random sample of working telephones, but even here complications arise. Different prefixes correspond to different numbers of working phones, and perhaps to different mixes of residential and business phones. Further, the probability that a number is answered may depend on the economic status of the owner. An important part of econometric analysis is determining when deviations from simple random sampling matter, and developing methods for dealing with them.

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 selects geographical units (e.g., cities, census tracts, telephone prefixes), 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 is the Liberty telephone poll in 1936 that predicted that Roosevelt would lose the Presidential election that he won in a landslide. The problem was that only the rich had telephones in 1936, so the sample was systematically biased. 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. Claims that a given sample frame and protocol have produced a simple random sample also deserve scrutiny,

An arena where the sampling theory is particularly obscure is in analysis of economic time-series. Here, one is observing a slice of history, and the questions are what is the population is from which this sample is drawn, and in what sense does this slice has properties of a random sample. One way statisticians have thought about this is to visualize our universe as being one draw from a population of "parallel universes".. This helps for doing formal probability theory, but is unsatisfying for the economist whose target is a hypothesis about the one universe we are in. 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 turn in later chapters to the problems that arise under alternative sampling protocols.

5.3. STATISTICAL DECISIONS

5.3.1. The process of statistical estimation can be thought of as decision-making under uncertainty. The economic problem faced by Cab Franc 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 $\mathbf{x}$ 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 $f(\mathbf{x},\theta)$. The density f is called the *likelihood* function of the data. The unknown parameter θ_o might be some population feature, such as age of retirement in the example discussed at the beginning of this chapter. 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 in statistical decisions? 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 avoid mistakes. For the example of retirement behavior, making a mistake on the retirement age may cause the planned medicare budget to go out of balance, and cost may be a known function of the magnitude of the unanticipated imbalance. However, if there are multiple clients, or the analysis is being performed for the scientific audience, there may not be precise costs, and it may be necessary to provide sufficient information from the analysis so that potential users can determine their most appropriate action based on their personal cost assessments. Before considering this situation, we will look at the case where there is a known cost function $C(\theta,\theta_o,\mathbf{x})$ that depends on the true parameter value θ_o and on the inference θ made from the data, and in general can also depend directly on $\mathbf{x}$.

5.3.2. A decision rule, or *action*, will be a mapping $T(\cdot)$ from the data $\mathbf{x}$ into the space of possible θ values. Note that while $T(\mathbf{x})$ depends on the data $\mathbf{x}$, it cannot depend directly on the unknown parameter θ_o, only indirectly through the influence of θ_o on the determination of $\mathbf{x}$. Because the data are random variables, $T(\cdot)$ is also a random variable, and it will have a density $\psi(t,\theta_o)$ that could be obtained from $f(\mathbf{x},\theta_o)$ by considering a one-to-one transformation from $\mathbf{x}$ to a vector that contains $T(\mathbf{x})$ and is filled out with some additional variables $\mathbf{Z}(\mathbf{x})$. The cost associated with the action $T(\cdot)$, given data $\mathbf{x}$, is $C(T(\mathbf{x}),\theta_o,\mathbf{x})$. One would like to choose this to be as small as possible, but the problem is that usually one cannot do this without knowing θ_o. However, the client may, prior to the observation of $\mathbf{x}$, 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 prior beliefs, 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 $\mathbf{x}$ 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 $\mathbf{x}$. Using this posterior density, the expected cost for an action $T(\mathbf{x})$ is

$$(2) \qquad R(T(\mathbf{x}),\mathbf{x}) = \int_{-\infty}^{+\infty} C(T(\mathbf{x}),\theta_o,\mathbf{x})p(\theta_o|\mathbf{x})d\theta_o$$

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

This expected cost is called the *Bayes risk*. It depends on the function $T(\cdot)$. The optimal action $T^*(\cdot)$ is the function $T(\cdot)$ that minimizes the expected cost for each $\mathbf{x}$, and therefore minimizes the Bayes risk. One has $R(T^*(x),\mathbf{x}) \le R(T^*(x) + \lambda,\mathbf{x})$ for a scalar λ, implying for each x the first-order condition $0 = \int_{-\infty}^{+\infty} \nabla_t C(T(\mathbf{x}),\theta_o,\mathbf{x})p(\theta_o|\mathbf{x})d\theta_o$. A strategy $T'(\mathbf{x})$ is called *inadmissible* if there is a second strategy $T''(\mathbf{x})$ such that $C(T''(\mathbf{x}),\theta,\mathbf{x}) \le C(T'(\mathbf{x}),\theta,\mathbf{x})$ for all θ for which $f(\mathbf{x},\theta) > 0$, with the inequality strict for some θ. Clearly the search for the optimal action $T^*(\mathbf{x})$ can be confined to the set of strategies that are admissible. In general, it is not obvious what the optimal action $T^*(\mathbf{x})$ that solves this problem looks like. A few examples help to provide some intuition:

(I) Suppose $C(\theta,\theta_o,\mathbf{x}) = (\theta - \theta_o)^2$, a quadratic cost function in which cost is proportional to the square of the distance of the estimator $T(\mathbf{x})$ from the true value θ_o. For a given $\mathbf{x}$, the argument $\theta = T(\mathbf{x})$ that minimizes (2) has to satisfy the first- order condition

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

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

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

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

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

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 $C(\theta,\theta_o,\mathbf{x}) = -1/2\alpha$ for $|\theta-\theta_o| \le \alpha$, and $C(\theta,\theta_o,\mathbf{x}) = 0$ 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; with α a positive parameter. The criterion (2) requires that $\theta = T^*(\mathbf{x})$ be chosen to minimize the expression $(-1/2\alpha)\int_{\theta-\alpha}^{\theta+\alpha} p(\theta_o|\mathbf{x})\cdot d\theta_o$. If α is very small, then $(-1/2\alpha)\int_{\theta-\alpha}^{\theta+\alpha} p(\theta_o|\mathbf{x})\cdot d\theta_o \simeq$ $-p(\theta|\mathbf{x})$. 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 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) $$0 = \frac{\nabla_\theta f(\boldsymbol{x},\theta)}{f(\boldsymbol{x},\theta)} + \frac{\nabla_\theta h(\theta)}{h(\theta)} .$$

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.

5.3.3. 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 only 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 "40 percent of workers age 65 will work another year if medicare is unavailable", 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 an event occurs, even if the "event" is something like a mathematical theorem whose truth is completely determinable by logic, and not the result of some cosmic coin toss. (In this case, if you do not know if the theorem is true, your probability may reflect your opinion of the mathematical abilities of its author.) How one thinks about probabilities influences how one thinks about an economic hypothesis, such as the hypothesis that retirement age does not depend on the age of medicare eligibility. In classical statistics, a 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 more meaningful to talk about a high or low probability of the hypothesis being true.

5.3.4. 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. (Bayesian computations can be quite difficult, and this is a second major barrier.) There are three possible options in the situation where there is not a well-defined prior.

5.3.5. The first 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" priors may not be proper probability densities that integrate to one.

This problem can be skirted 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".

5.3.6. The second option is based on the idea that you are in a game against Nature in which Nature plays θ_o and reveals information **x** about her strategy, and you know that **x** is a draw from the DGP $f(\mathbf{x},\theta_o)$. You then play $T(\mathbf{x})$. Of course, if you had a prior $h(\theta_o)$, which in this context might be interpreted as a conjecture about Nature's play, you could adopt the optimal Bayes strategy $T^*(x)$. A conservative strategy in games is to play in such a way that you minimize the maximum cost your opponent can impose on you. This strategy picks $T^*(\mathbf{x}) = \text{argmin}_t \max_h R(t,\mathbf{x},h)$, where R is the Bayes risk, now written with an argument h to emphasize that it depends on the prior h. The idea is that the worst Nature can do is draw θ_o from the prior that is the least favorable to you in terms of costs, and this strategy minimizes this maximum expected cost. This is called a *minimax* strategy. Unless the problem has some compactness properties, the minimax strategy may not exist, although there may be a sequence of strategies that come close. A minimax strategy 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. There may however “meta-Bayesian” solutions in which search for a least favorable prior is limited to a class that the analyst considers “possible”.

5.3.7. The final option 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. Suppose there is a one-to-one transformation of the data **x** into two components (**y**,**z**) so that the likelihood function $f(\mathbf{x},\theta)$ factors into the product of a marginal density of **y** that depends on θ, and a conditional density of **z**, given **y**, that does not depend on θ, $f(\mathbf{x},\theta) \equiv f_1(\mathbf{y},\theta)\cdot f_2(\mathbf{z}|\mathbf{y})$. In this case, **y** is termed a *sufficient statistic* for θ. If one forms the posterior density of θ given **x** and a prior density h, one has in this case

$$(5) \qquad p(\theta|\mathbf{x}) = \frac{f_1(y,\theta)\cdot f_2(z|y)\cdot h(\theta)}{\int_{-\infty}^{+\infty} f_1(y,\theta')\cdot f_2(z|y)\cdot h(\theta')d\theta'} = \frac{f_1(y,\theta)\cdot h(\theta)}{\int_{-\infty}^{+\infty} f_1(y,\theta')\cdot h(\theta')d\theta'} = p(\theta|\mathbf{y}).$$

Then, all the information to be learned about θ from **x**, reflected in the posterior density, can be learned from the summary data **y**. Then it is unnecessary to retain all the original data in **x** for purposes of statistical inference on θ; rather it is enough to retain the sufficient statistic **y**. By reporting **y**, the econometrician leaves the user completely free to form a prior, and calculate the posterior likelihood and the action that minimizes the user's Bayes risk. 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.

5.4. 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 consensus 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 alternative to the Bayesian statistical decision-making framework. We will do this in the next chapter.

5.5. EXERCISES

1. Suppose you are concerned about the question of whether skilled immigrants make a positive net contribution to the U.S. economy, to the non-immigrant residents, and to the domestic workers who are competing directly for the same skilled jobs. If you could design an experiment to measure these effects, how would you do it? If you have to rely on a natural experiment, what would you look for?

2. Suppose you have a random sample of size n from a population with CDF F(x) which has mean μ and variance σ^2. You estimate μ by forming the sample average, or mean of the empirical distribution F_n. The distribution of this estimator could be determined by drawing repeated samples from F, forming the empirical distribution of the sample averages, and taking the limit. An approximation to this calculation, called the *bootstrap*, starts from the known empirical distribution of the original sample F_n rather than the unknown F. Try this computationally. Take F to be uniform on [0,1], and draw a base sample of size 10. Now estimate the CDF of the sample mean by (1) repeatedly

sampling from the uniform distribution and (2) sampling with replacement from the base sample. Do 100 draws from each, and compare their medians.

3. Discuss how you would go about drawing a simple random sample of commuters in the San Francisco Bay Area. What problems do you face in defining the universe and the sample frame. Discuss ways in which you could implement the sampling. What problems are you likely to encounter?

4. Review the decision problem of Cab Franc in Chapter 1, and put it in the terminology of Section 5.3.2. Discuss the impact on Cab's decision of the particular loss function he has.