depends on the whole sampling distribution, and is concerned about long-run average performance of the rule over repeated experiments, is affected. A couple of famous examples of standard inferential approaches that violate the Likelihood Principle in somewhat embarrassing ways are in Problems 7.12 and 8.5. ⋆

7.2.4 Nuisance parameters

The realistic specification of a sampling model often requires parameters other than those of primary interest. These additional parameters are called "nuisance parameters." This is one of the very few reasonably named concepts in statistics, as it causes all kinds of trouble to frequentist and likelihood theories alike. Basu (1975) gives a critical discussion.

From a decision-theoretic standpoint, we can think of nuisance parameters as those which appear in the sampling distribution, but not in the loss function. We formalize this notion from a decision-theoretic viewpoint and establish a general result for dealing with nuisance parameters in statistics. The bottom line is that the expected utility principle justifies averaging the likelihood and the prior over the possible values of the nuisance parameter and taking things from there. In probabilistic terminology, nuisance parameters can be integrated out, and the original decision problem can be replaced by its marginal version. More specifically:

Theorem 7.1 *If θ can be partitioned into (θ^*, η) such that the loss $L(\theta, a)$ depends on θ only through θ^*, then η is a nuisance parameter and the Bayes rule for the problem with likelihood $f(x|\theta)$ and prior $\pi(\theta)$ is the same as the Bayes rule for the problem with likelihood*

$$f^*(x|\theta^*) = \int_H f(x|\theta)\pi(\eta|\theta^*)d\eta$$

and prior

$$\pi(\theta^*) = \int_H \pi(\theta)d\eta$$

where H is the domain of η.

Proof: We assume that all integrals involved are finite. Take any decision rule δ. The Bayes risk is

$$\begin{aligned} r(\pi, \delta) &= \int_\Theta \int_{\mathcal{X}} L(\theta, \delta) f(x|\theta)\pi(\theta)dxd\theta \\ &= \int_{\mathcal{X}} \int_{\Theta^*} L(\theta^*, \delta) \int_H f(x|\theta)\pi(\theta)d\eta d\theta^* dx \\ &= \int_{\mathcal{X}} \int_{\Theta^*} L(\theta^*, \delta) \int_H f(x|\theta)\pi(\eta|\theta^*)d\eta\pi(\theta^*)d\theta^* dx \\ &= \int_{\mathcal{X}} \int_{\Theta^*} L(\theta^*, \delta) f^*(x|\theta^*)\pi(\theta^*)d\theta^* dx, \end{aligned}$$

that is the Bayes risk for the problem with likelihood $f^*(x|\theta^*)$ and prior $\pi(\theta^*)$. We used the independence of L on η, and the relation $\pi(\theta) = \pi(\eta, \theta^*) = \pi(\eta|\theta^*)\pi(\theta^*)$. □

This theorem puts to rest the issue of nuisance parameters in every conceivable statistical problem, as long as one can specify reasonable priors, and compute integrals. Neither is easy, of course. Priors on high-dimensional nuisance parameters can be very difficult to assess based on expert knowledge and often include surprises in the form of difficult-to-anticipate implications when nuisance parameters are integrated out. Integration in high dimension has made much progress over the last 20 years, thanks mostly to Markov chain Monte Carlo (MCMC) methods (Robert & Casella 1999), but is still hard, in part because we tend to adapt to this progress and specify models that are at the limit of what is computable. Nonetheless, the elegance and generality of the solution are compelling.

Another way to interpret the Bayesian solution to the nuisance parameter problem is to look at posterior expected losses. An argument similar to that used in the proof of Theorem 7.1 would show that one can equivalently compute the posterior expected losses based on the marginal posterior distribution of θ^* given by

$$\pi(\theta^*|x) = \int_H \pi(\theta, \eta|x)d\eta = \int_H \pi(\theta|x, \eta)\pi(\eta|x)d\eta.$$

This highlights the fact that the Bayes rule is potentially affected by any of the features of the posterior distribution $\pi(\eta|x)$ of the nuisance parameter, including all aspects of the uncertainty that remains about them after observing the data. This is in contrast to approaches that eliminate nuisance parameters by "plugging in" best estimates either in the likelihood function or in the decision rule itself. The empirical Bayes approach of Section 9.2.2 is an example.

7.3 The travel insurance example

In this section we introduce a mildly realistic medical example that will hopefully afford the simplest possible illustration of the concepts introduced in this chapter and also give us the excuse to introduce some terminology and graphics from decision analysis. We will return to this example when we consider multistage decision problems in Chapters 12 and 13.

Suppose that you are from the United States and are about to take a trip overseas. You are not sure about the status of your vaccination against a certain mild disease that is common in the country you plan to visit, and need to decide whether to buy medical insurance for the trip. We will assume that you will be exposed to the disease, but you are uncertain about whether your present immunization will work. Based on aggregate data on western tourists, the chance of developing the disease during the trip is about 3% overall. Treatment and hospital abroad would normally cost you, say, 1000 dollars. There is also a definite loss in quality of life in going all the way to an exotic country and being grounded at a local hospital instead of making the most

out of your experience, but we are going to ignore this aspect here. On the other hand, if you buy a travel insurance plan, which you can do for 50 dollars, all your expenses will be covered. This is a classical gamble versus sure outcome situation. Table 7.6 summarizes the loss function for this problem.

For later reference we are going to represent this simple case using a decision tree. In a decision tree, a square denotes a *decision node* or *decision point*. The decision maker has to decide among actions, represented by branches stemming out from the decision node. A circle represents a *chance node* or *chance point*. Each branch out of the circle represents, in this case, a state of nature, though circles could also be used for experimental results. On the right side of the decision tree we have the consequences. Figure 7.3 shows the decision tree for our problem.

In a Bayesian mode, you use the expected losses to evaluate the two actions, as follows:

$$\text{No insurance:} \quad \text{Expected loss} = 1000 \times 0.03 + 0 \times 0.97 = 30$$

$$\text{Insurance:} \quad \text{Expected loss} = 50 \times 0.03 + 50 \times 0.97 = 50.$$

Table 7.6 Monetary losses associated with buying and with not buying travel insurance for the trip.

Actions	Events	
	θ_1: ill	θ_2: not ill
Insurance	50	50
No insurance	1000	0

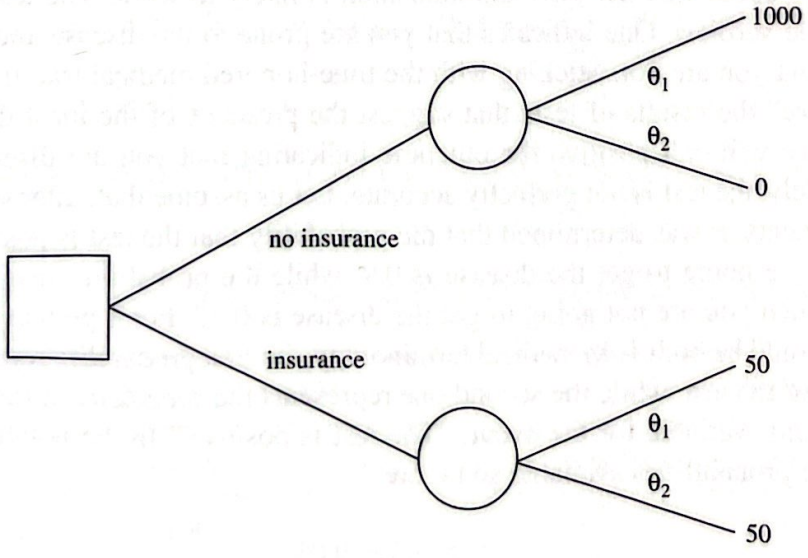

Figure 7.3 Decision tree for the travel insurance example. This is a single-stage tree, because it includes only one decision node along any given path.

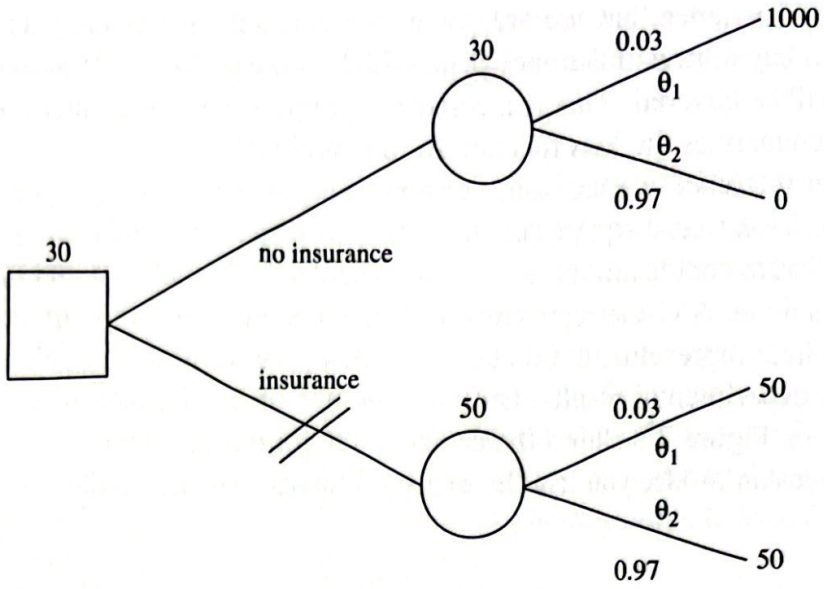

Figure 7.4 Solved decision tree for the medical insurance example. At the top of each chance node we have the expected loss, while at the top of the decision node we have the minimum expected loss. Alongside the branches stemming out from the chance node we have the probabilities of the states of nature. The action that is not optimal is crossed out by a double line.

The Bayes decision is the decision that minimizes the expected loss—in this case not to buy the insurance. However, if the chance of developing the disease was 5% or greater, the best decision would be to buy the insurance. The solution to this decision problem is represented in Figure 7.4.

You can improve your decision making by gathering data on how likely you are to get the disease. Imagine you have the option of undergoing a medical test that informs you about whether your immunization is likely to work. The test has only two possible verdicts. One indicates that you are prone to the disease and the other indicates that you are not. Sticking with the time-honored medical tradition of calling "positive" the results of tests that suggest the presence of the most devastating illnesses, we will call positive the outcome indicating that you are disease prone. Unfortunately, the test is not perfectly accurate. Let us assume that, after some clinical experiments, it was determined that the probability that the test is positive when you really are going to get the disease is 0.9, while the probability that the test is negative when you are not going to get the disease is 0.77. For a perfect test, these numbers would be both 1. In medical terminology, the first probability represents the *sensitivity* of the test, while the second one represents the *specificity* of the test. Call x the indicator variable for the event, "The test is positive." In the notation of this chapter, the probabilities available so far are

$$\pi(\theta_1) = 0.03$$
$$f(x = 1|\theta_1) = 0.90$$
$$f(x = 0|\theta_2) = 0.77.$$

After the test, your individual chances of illness will be different from the overall 3%. The test could provide valuable information and potentially alter your chosen course of action. The question of this chapter is precisely how to use the results of the test to make a better decision. The test seems reliable enough that we may want to buy the insurance if the test is positive and not otherwise. Is this right?

To answer this question, we will consider decision rules. In our example there are two possible experimental outcomes and two possible actions so there are a total of four possible decisions rules. These are

$\delta_0(x)$: Do not buy the insurance.
$\delta_1(x)$: Buy the insurance if $x = 1$. Otherwise, do not.
$\delta_2(x)$: Buy the insurance if $x = 0$. Otherwise, do not.
$\delta_3(x)$: Buy the insurance.

Decision rules δ_0 and δ_3 choose the same action irrespective of the outcome of the test: they are constant functions. Decision rule δ_1 does what comes naturally: buy the insurance only if the test indicates that you are disease prone. Decision rule δ_2 does exactly the opposite. As you might expect, it will not turn out to be very competitive.

Let us now look at the losses associated with each decision rule ignoring, for now, any costs associated with testing. Of course, the loss for rules δ_1 and δ_2 now depends on the data. We can summarize the situation as shown in Table 7.7.

Two unknowns will affect how good our choice will turn out to be: the test result and whether you will be ill during the trip. As in equation (7.14) we can choose the Bayes rule by averaging out both, beginning with averaging losses by state, and then further averaging the results to obtain overall average losses. The results are shown in Table 7.8. To illustrate how entries are calculated, consider δ_1. The average risk if $\theta = \theta_1$ is

$$1000f(x = 0|\theta_1) + 50f(x = 1|\theta_1) = 1000 \times 0.10 + 50 \times 0.90 = 145.0$$

while the average risk if $\theta = \theta_2$ is

$$0f(x = 0|\theta_2) + 50f(x = 1|\theta_2) = 0 \times 0.77 + 50 \times 0.23 = 11.5,$$

Table 7.7 Loss table for the decision rules in the travel insurance example.

	θ_1: ill		θ_2: not ill	
	$x = 0$	$x = 1$	$x = 0$	$x = 1$
$\delta_0(x)$	\$1000	\$1000	\$0	\$0
$\delta_1(x)$	\$1000	\$50	\$0	\$50
$\delta_2(x)$	\$50	\$1000	\$50	\$0
$\delta_3(x)$	\$50	\$50	\$50	\$50

Table 7.8 Average losses by state and overall for the decision rules in the travel insurance example.

	Average losses by state		Average losses overall
	θ_1: ill	θ_2: not ill	
$\delta_0(x)$	\$1000.0	\$0.0	\$30.0
$\delta_1(x)$	\$145.0	\$11.5	\$15.5
$\delta_2(x)$	\$905.0	\$38.5	\$64.5
$\delta_3(x)$	\$50.0	\$50.0	\$50.0

so that the overall average is

$$145.0 \times \pi(\theta_1) + 11.5 \times \pi(\theta_2) = 15.5 = 145.0 \times 0.03 + 11.5 \times 0.97 = 15.5.$$

Strategy $\delta_1(x)$ is the Bayes strategy as it minimizes the overall expected loss. This calculation is effectively considering the losses in Table 7.7 and computing the expectation of each row with respect to the joint distribution of θ and x. In this sense it is consistent with preferences expressed prior to observing x. You are bound to stick to the optimal rule after you actually observe x only if you also agree with the before/after axiom of Section 5.2.3. Then, an alternative derivation of the Bayes rule could have been worked out directly by computing posterior expected losses given $x = 1$ and $x = 0$, as we know from Section 7.2.2 and equation (7.15).

So far, in solving the decision problem we utilized the Bayes principle. Alternatively, if you follow the minimax principle, your goal is avoiding the largest possible loss. Let us start with the case in which no data are available. In our example, the larges loss is 50 dollars if you buy the insurance and 1000 if you do not. By this principle you should buy the medical insurance. In fact, as we examine Table 7.6, we note that the greatest loss is associated with event θ_1 no matter what the action is. Therefore the maximization step in the minimax calculation will resolve the uncertainty about θ by assuming, pessimistically, that you will become ill, no matter how much evidence you may accumulate to the contrary. To alleviate this drastic pessimism, let us express the losses in "regret" form. The argument is as follows. If you condition on getting ill, the best you can do is a loss of \$50, by buying the medical insurance. The alternative action entails a loss of \$1000. When you assess the worthiness of this action, you should compare the loss to the best (smallest) loss that you could have obtained. You do indeed lose \$1000, but your "regret" is only for the \$950 that you could have avoided spending. Applying equation (7.2) to Table 7.6 gives Table 7.9.

When reformulating the decision problem in terms of regret, the expected losses are

$$\begin{aligned} \text{No insurance:} \quad & \text{Expected loss} = 950 \times 0.03 + 0 \times 0.97 = 28.5 \\ \text{Insurance:} \quad & \text{Expected loss} = 0 \times 0.03 + 50 \times 0.97 = 48.5. \end{aligned}$$

Table 7.9 Regret loss table for the actions in the travel insurance example.

		Event	
		θ_1	θ_2
Decision:	insurance	\$0	\$50
	no insurance	\$950	\$0

Table 7.10 Risk table for the decision rules in the medical insurance example when using regret losses.

	Risk $R(\theta, \delta)$ by state		Largest risk	Average risk $r(\pi, \delta)$
	θ_1	θ_2		
$\delta_0(x)$	\$950	\$0	\$950	\$28.5
$\delta_1(x)$	\$95	\$11.5	\$95	\$14.0
$\delta_2(x)$	\$855	\$38.5	\$855	\$63.0
$\delta_3(x)$	\$0	\$50	\$50	\$48.5

The Bayes action remains the same. The expected losses become smaller, but the expected loss of every action becomes smaller by the same amount. On the contrary, the minimax action may change, though in this example it does not. The minimax solution is still to buy the insurance.

Does the optimal minimax decision change depending on the test results? In Table 7.10 we derive the Bayes and minimax rules using the regret losses. Strategy δ_1 is the Bayes strategy as we had seen before. We also note that δ_2 is dominated by δ_1, that is it has a higher risk than δ_1 irrespective of the true state of the world. Using the minimax approach, the optimal decision is δ_3, that is it is still optimal to buy the insurance irrespective of the test result. This conclusion depends on the losses, sensitivity, and specificity, and different rules could be minimax if these parameters were changed.

This example will reappear in Chapters 12 and 13, when we will consider both the decision of whether to do the test and the decision of what to do with the information.