7.1 Introduction

Until now, we have dealt only with the making of decisions or inferences. Another very important aspect of statistics is that of the choice of experiment, commonly called experimental design. Being as this choice must (usually) be made before the data (and hence the posterior distribution) can be obtained, the subject is frequently called preposterior analysis by Bayesians.

The goal of preposterior analysis is to choose the experiment or experimental design which minimizes overall cost. This overall cost consists of the decision loss and the cost of conducting and analyzing the experiment. We will essentially ignore the cost of analyzing the results (after all, statisticians are so underpaid that the cost of their labors is usually negligible), and thus consider only the decision loss and the cost of experimentation. Note that these last quantities are in opposition to each other. To lower the decision loss it will generally be necessary to run a larger experiment, whereby the experimental cost will be increased. In this chapter, we will be concerned with the balancing of these two costs.

The general problem of experimental design is very complicated, concerning choices among quite different experiments (say, a choice between a completely randomized design and a randomized block design, or maybe a choice of the values of the independent variables in a regression study). Though of considerable interest, investigation of such general questions is beyond the scope of this book. We will content ourselves with a study of the simplest design problem, that of deciding when to stop sampling.

To get specific, assume random variables $X_1, X_2, \ldots$ are available for observation. Let $\mathcal{X}_i$ be the sample space of $X_i$, define $X^j = (X_1, X_2, \ldots, X_j)$, and assume that $X^j$ has density $f_j(x^j|\theta)$ on $\mathcal{X}^j = \mathcal{X}_1 \times \cdots \times \mathcal{X}_j$. As usual,
\( \theta \in \Theta \) is the unknown state of nature, concerning which some inference or decision is to be made. Most of the examples we will consider deal with situations in which the \( X_i \) are independent observations from a common density \( f(x|\theta) \). In such a situation,

\[
f_j(x^j|\theta) = \prod_{i=1}^{j} f(x_i|\theta),
\]

and we will say that \( X_1, X_2, \ldots \) is a sequential sample from the density \( f(x|\theta) \). It will be assumed that the observations can be taken in stages, or sequentially. This means that after observing, say, \( X^j = (X_1, \ldots, X_j) \), the experimenter has the option of either making an immediate decision or taking further observations.

The experimental cost in this setup is simply the cost of taking observations. This cost can depend on many factors, two of the most crucial being the number of observations ultimately taken, and the way in which the observations are taken (i.e., one at a time, in groups, etc.). To quantify this, let \( n \) denote the number of observations ultimately taken, let \( s \) denote the manner in which the observations are taken, and, as usual, let \( a \in \mathcal{A} \) denote the action taken. Then

\[
L(\theta, a, n, s)
\]

will denote the overall loss or cost when \( \theta \) turns out to be the true state of nature.

Frequently, it will be the case that \( L(\theta, a, n, s) \) can be considered to be the sum of the decision loss, \( L(\theta, a) \), and the cost of observation (or sampling cost), \( C(n, s) \). This will happen when the decision maker has a (nearly) linear utility function, so that the combined loss is just the sum of the individual losses. If the utility function \( U \) is nonlinear and the situation involves, say, monetary gain or loss, then it will generally happen, instead, that

\[
L(\theta, a, n, s) = -U(G(\theta, a) - C(n, s)),
\]

where \( G(\theta, a) \) represents the monetary gain when the pair \((\theta, a)\) occurs and \( C(n, s) \) is the sampling cost.

Unfortunately, even the sequential decision problem in this generality is too hard to handle. The difficulty occurs in trying to deal with all possible methods, \( s \), of taking observations. We will, therefore, restrict ourselves to studying the two most common methods of taking observations. The first is the fixed sample size method, in which one preselects a sample size \( n \), observes \( X^n = (X_1, \ldots, X_n) \), and makes a decision. The overall loss for this situation will be denoted \( L^F(\theta, a, n) \). This problem will be considered in the next section.

The second common method of taking observations is that of sequential analysis, in which the observations are taken one at a time, with a decision being made, after each observation, to either cease sampling (and choose an action \( a \in \mathcal{A} \)) or take another observation. This is the situation that will be discussed in the bulk of the chapter (Sections 7.3 through 7.7), and so, for simplicity, the overall loss will just be written \( L(\theta, a, n) \) in this case.