James O. Berger

Statistical Decision Theory and Bayesian Analysis

Second Edition

With 23 Illustrations

Springer-Verlag
New York Berlin Heidelberg London Paris
Tokyo Hong Kong Barcelona Budapest

Contents

CHAPTER 1
Basic Concepts

1.1. Introduction

Decision theory, as the name implies, is concerned with the problem of making decisions. Statistical decision theory is concerned with the making of decisions in the presence of statistical knowledge which sheds light on some of the uncertainties involved in the decision problem. We will, for the most part, assume that these uncertainties can be considered to be unknown numerical quantities, and will represent them by θ (possibly a vector or matrix).

As an example, consider the situation of a drug company deciding whether or not to market a new pain reliever. Two of the many factors affecting its decision are the proportion of people for which the drug will prove effective (θ_1), and the proportion of the market the drug will capture (θ_2). Both θ_1 and θ_2 will be generally unknown, though typically experiments can be conducted to obtain statistical information about them. This problem is one of decision theory in that the ultimate purpose is to decide whether or not to market the drug, how much to market, what price to charge, etc.

Classical statistics is directed towards the use of sample information (the data arising from the statistical investigation) in making inferences about θ. These classical inferences are, for the most part, made without regard to the use to which they are to be put. In decision theory, on the other hand, an attempt is made to combine the sample information with other relevant aspects of the problem in order to make the best decision.

In addition to the sample information, two other types of information are typically relevant. The first is a knowledge of the possible consequences of the decisions. Often this knowledge can be quantified by determining the loss that would be incurred for each possible decision and for the various

possible values of θ. (Statisticians seem to be pessimistic creatures who think in terms of losses. Decision theorists in economics and business talk instead in terms of gains (utility). As our orientation will be mainly statistical, we will use the loss function terminology. Note that a gain is just a negative loss, so there is no real difference between the two approaches.)

The incorporation of a loss function into statistical analysis was first studied extensively by Abraham Wald; see Wald (1950), which also reviews earlier work in decision theory.

In the drug example, the losses involved in deciding whether or not to market the drug will be complicated functions of θ_1, θ_2, and many other factors. A somewhat simpler situation to consider is that of estimating θ_1, for use, say, in an advertising campaign. The loss in underestimating θ_1 arises from making the product appear worse than it really is (adversely affecting sales), while the loss in overestimating θ_1 would be based on the risks of possible penalties for misleading advertising.

The second source of nonsample information that is useful to consider is called prior information. This is information about θ arising from sources other than the statistical investigation. Generally, prior information comes from past experience about similar situations involving similar θ. In the drug example, for instance, there is probably a great deal of information available about θ_1 and θ_2 from different but similar pain relievers.

A compelling example of the possible importance of prior information was given by L. J. Savage (1961). He considered the following three statistical experiments:

1. A lady, who adds milk to her tea, claims to be able to tell whether the tea or the milk was poured into the cup first. In all of ten trials conducted to test this, she correctly determines which was poured first.
2. A music expert claims to be able to distinguish a page of Haydn score from a page of Mozart score. In ten trials conducted to test this, he makes a correct determination each time.
3. A drunken friend says he can predict the outcome of a flip of a fair coin. In ten trials conducted to test this, he is correct each time.

In all three situations, the unknown quantity θ is the probability of the person answering correctly. A classical significance test of the various claims would consider the null hypothesis (H_0) that $\theta = 0.5$ (i.e., the person is guessing). In all three situations this hypothesis would be rejected with a (one-tailed) significance level of 2^{-10}. Thus the above experiments give strong evidence that the various claims are valid.

In situation 2 we would have no reason to doubt this conclusion. (The outcome is quite plausible with respect to our prior beliefs.) In situation 3, however, our prior opinion that this prediction is impossible (barring a belief in extrasensory perception) would tend to cause us to ignore the experimental evidence as being a lucky streak. In situation 1 it is not quite clear what to think, and different people will draw different conclusions

according to their prior beliefs of the plausibility of the claim. In these three identical statistical situations, prior information clearly cannot be ignored.

The approach to statistics which formally seeks to utilize prior information is called Bayesian analysis (named after Bayes (1763)). Bayesian analysis and decision theory go rather naturally together, partly because of their common goal of utilizing nonexperimental sources of information, and partly because of some deep theoretical ties; thus, we will emphasize Bayesian decision theory in the book. There exist, however, an extensively developed non-Bayes decision theory and an extensively developed non-decision-theoretic Bayesian viewpoint, both of which we will also cover in reasonable depth.

1.2. Basic Elements

The unknown quantity θ which affects the decision process is commonly called the *state of nature*. In making decisions it is clearly important to consider what the possible states of nature are. The symbol Θ will be used to denote the set of all possible states of nature. Typically, when experiments are performed to obtain information about θ, the experiments are designed so that the observations are distributed according to some probability distribution which has θ as an unknown parameter. In such situations θ will be called the *parameter* and Θ the *parameter space*.

Decisions are more commonly called *actions* in the literature. Particular actions will be denoted by a, while the set of all possible actions under consideration will be denoted $\mathcal{A}$.

As mentioned in the introduction, a key element of decision theory is the loss function. If a particular action a_1 is taken and θ_1 turns out to be the true state of nature, then a loss $L(\theta_1, a_1)$ will be incurred. Thus we will assume a *loss function* $L(\theta, a)$ is defined for all $(\theta, a) \in \Theta \times \mathcal{A}$. For technical convenience, only loss functions satisfying $L(\theta, a) \geq -K > -\infty$ will be considered. This condition is satisfied by all loss functions of interest. Chapter 2 will be concerned with showing why a loss function will typically exist in a decision problem, and with indicating how a loss function can be determined.

When a statistical investigation is performed to obtain information about θ, the outcome (a random variable) will be denoted X. Often X will be a vector, as when $X = (X_1, X_2, \ldots, X_n)$, the X_i being independent observations from a common distribution. (From now on vectors will appear in boldface type; thus $\mathbf{X}$.) A particular realization of X will be denoted x. The set of possible outcomes is the *sample space*, and will be denoted $\mathcal{X}$. (Usually $\mathcal{X}$ will be a subset of R^n, n-dimensional Euclidean space.)

The probability distribution of X will, of course, depend upon the unknown state of nature θ. Let $P_\theta(A)$ or $P_\theta(X \in A)$ denote the probability

of the event $A (A \subset \mathscr{X})$, when θ is the true state of nature. For simplicity, X will be assumed to be either a continuous or a discrete random variable, with density $f(x|\theta)$. Thus if X is continuous (i.e., has a density with respect to Lebesgue measure), then

$$P_\theta(A) = \int_A f(x|\theta)dx,$$

while if X is discrete, then

$$P_\theta(A) = \sum_{x \in A} f(x|\theta).$$

Certain common probability densities and their relevant properties are given in Appendix 1.

It will frequently be necessary to consider expectations over random variables. The expectation (over X) of a function $h(x)$, for a given value of θ, is defined to be

$$E_\theta[h(X)] = \begin{cases} \int_{\mathscr{X}} h(x)f(x|\theta)dx & \text{(continuous case)}, \\ \sum_{x \in \mathscr{X}} h(x)f(x|\theta) & \text{(discrete case)}. \end{cases}$$

It would be cumbersome to have to deal separately with these two different expressions for $E_\theta[h(X)]$. Therefore, as a convenience, we will define

$$E_\theta[h(X)] = \int_{\mathscr{X}} h(x)dF^X(x|\theta),$$

where the right-hand side is to be interpreted as in the earlier expression for $E_\theta[h(X)]$. (This integral can, of course, be considered a Riemann-Stieltjes integral, where $F^X(x|\theta)$ is the cumulative distribution function of X. Readers not familiar with such terms can just treat the integral as a notational device.) Note that, in the same way, we can write

$$P_\theta(A) = \int_A dF^X(x|\theta).$$

Frequently, it will be necessary to clarify the random variables over which an expectation or probability is being taken. Superscripts on E or P will serve this role. (A superscript could be the random variable, its density, its distribution function, or its probability measure, whichever is more convenient.) Subscripts on E will denote parameter values at which the expectation is to be taken. When obvious, subscripts or superscripts will be omitted.

The third type of information discussed in the introduction was prior information concerning θ. A useful way of talking about prior information is in terms of a probability distribution on Θ. (Prior information about θ is seldom very precise. Therefore, it is rather natural to state prior beliefs

in terms of probabilities of various possible values of θ being true.) The symbol $\pi(\theta)$ will be used to represent a prior density of θ (again for either the continuous or discrete case). Thus if $A \subset \Theta$,

$$P(\theta \in A) = \int_A dF^{\pi}(\theta) = \begin{cases} \int_A \pi(\theta)\, d\theta & \text{(continuous case)}, \\ \sum_{\theta \in A} \pi(\theta) & \text{(discrete case)}. \end{cases}$$

Chapter 3 discusses the construction of prior probability distributions, and also indicates what is meant by probabilities concerning θ. (After all, in most situations there is nothing "random" about θ. A typical example is when θ is an unknown but fixed physical constant (say the speed of light) which is to be determined. The basic idea is that probability statements concerning θ are then to be interpreted as "personal probabilities" reflecting the degree of personal belief in the likelihood of the given statement.)

Three examples of use of the above terminology follow.

EXAMPLE 1. In the drug example of the introduction, assume it is desired to estimate θ_2. Since θ_2 is a proportion, it is clear that $\Theta = \{\theta_2 : 0 \le \theta_2 \le 1\} = [0, 1]$. Since the goal is to estimate θ_2, the action taken will simply be the choice of a number as an estimate for θ_2. Hence $\mathcal{A} = [0, 1]$. (Usually $\mathcal{A} = \Theta$ for estimation problems.) The company might determine the loss function to be

$$L(\theta_2, a) = \begin{cases} \theta_2 - a & \text{if } \theta_2 - a \ge 0, \\ 2(a - \theta_2) & \text{if } \theta_2 - a \le 0. \end{cases}$$

(The loss is in units of "utility," a concept that will be discussed in Chapter 2.) Note that an overestimate of demand (and hence overproduction of the drug) is considered twice as costly as an underestimate of demand, and that otherwise the loss is linear in the error.

A reasonable experiment which could be performed to obtain sample information about θ_2 would be to conduct a sample survey. For example, assume n people are interviewed, and the number X who would buy the drug is observed. It might be reasonable to assume that X is $\mathcal{B}(n, \theta_2)$ (see Appendix 1), in which case the sample density is

$$f(x|\, \theta_2) = \binom{n}{x} \theta_2^x (1 - \theta_2)^{n-x}.$$

There could well be considerable prior information about θ_2, arising from previous introductions of new similar drugs into the market. Let's say that, in the past, new drugs tended to capture between $\frac{1}{10}$ and $\frac{1}{5}$ of the market, with all values between $\frac{1}{10}$ and $\frac{1}{5}$ being equally likely. This prior information could be modeled by giving θ_2 a $\mathcal{U}(0.1, 0.2)$ prior density, i.e., letting

$$\pi(\theta_2) = 10 I_{(0.1, 0.2)}(\theta_2).$$

The above development of L, f, and π is quite crude, and usually much more detailed constructions are required to obtain satisfactory results. The techniques for doing this will be developed as we proceed.

EXAMPLE 2. A shipment of transistors is received by a radio company. It is too expensive to check the performance of each transistor separately, so a sampling plan is used to check the shipment as a whole. A random sample of n transistors is chosen from the shipment and tested. Based upon X, the number of defective transistors in the sample, the shipment will be accepted or rejected. Thus there are two possible actions: a_1—accept the shipment, and a_2—reject the shipment. If n is small compared to the shipment size, X can be assumed to have a $\mathcal{B}(n, \theta)$ distribution, where θ is the proportion of defective transistors in the shipment.

The company determines that their loss function is $L(\theta, a_1) = 10\theta$, $L(\theta, a_2) = 1$. (When a_2 is decided (i.e., the lot is rejected), the loss is the constant value 1, which reflects costs due to inconvenience, delay, and testing of a replacement shipment. When a_1 is decided (i.e., the lot is accepted), the loss is deemed proportional to θ, since θ will also reflect the proportion of defective radios produced. The factor 10 indicates the relative costs involved in the two kinds of errors.)

The radio company has in the past received numerous other transistor shipments from the same supplying company. Hence they have a large store of data concerning the value of θ on past shipments. Indeed a statistical investigation of the past data reveals that θ was distributed according to a $\mathcal{B}e(0.05, 1)$ distribution. Hence

$$\pi(\theta) = (0.05)\theta^{-0.95} I_{[0,1]}(\theta).$$

EXAMPLE 3. An investor must decide whether or not to buy rather risky ZZZ bonds. If the investor buys the bonds, they can be redeemed at maturity for a net gain of \$500. There could, however, be a default on the bonds, in which case the original \$1000 investment would be lost. If the investor instead puts his money in a "safe" investment, he will be guaranteed a net gain of \$300 over the same time period. The investor estimates the probability of a default to be 0.1.

Here $\mathcal{A} = \{a_1, a_2\}$, where a_1 stands for buying the bonds and a_2 for not buying. Likewise $\Theta = \{\theta_1, \theta_2\}$, where θ_1 denotes the state of nature "no default occurs" and θ_2 the state "a default occurs." Recalling that a gain is represented by a negative loss, the loss function is given by the following table.

	a_1	a_2
θ_1	-500	-300
θ_2	1000	-300

(When both Θ and $\mathcal{A}$ are finite, the loss function is most easily represented by such a table, and is called a *loss matrix*. Actions are typically placed along the top of the table, and θ values along the side.) The prior information can be written as $\pi(\theta_1) = 0.9$ and $\pi(\theta_2) = 0.1$.

Note that in this example there is no sample information from an associated statistical experiment. Such a problem is called a *no-data* problem.

It should not be construed from the above examples that every problem will have a well-defined loss function and explicit prior information. In many problems these quantities will be very vague or even nonunique. The most important examples of this are problems of statistical inference. In statistical inference the goal is not to make an immediate decision, but is instead to provide a "summary" of the statistical evidence which a wide variety of future "users" of this evidence can easily incorporate into their own decision-making processes. Thus a physicist measuring the speed of light cannot reasonably be expected to know the losses that users of his result will have.

Because of this point, many statisticians use "statistical inference" as a shield to ward off consideration of losses and prior information. This is a mistake for several reasons. The first is that reports from statistical inferences should (ideally) be constructed so that they can be easily utilized in individual decision making. We will see that a number of classical inferences are failures in this regard.

A second reason for considering losses and prior information in inference is that the investigator may very well possess such information; he will often be very informed about the uses to which his inferences are likely to be put, and may have considerable prior knowledge about the situation. It is then almost imperative that he present such information in his analysis, although care should be taken to clearly separate "subjective" and "objective" information (but see Subsection 1.6.5 and Section 3.7).

The final reason for involvement of losses and prior information in inference is that choice of an inference (beyond mere data summarization) can be viewed as a decision problem, where the action space is the set of all possible inference statements and a loss function reflecting the success in conveying knowledge is used. Such "inference losses" will be discussed in Subsections 2.4.3 and 4.4.4. And, similarly, "inference priors" can be constructed (see Sections 3.3 and 4.3) and used to compelling advantage in inference.

While the above reasons justify specific incorporation of loss functions and prior information into inference, decision theory can be useful even when such incorporation is proscribed. This is because many standard inference criteria can be formally reproduced as decision-theoretic criteria with respect to certain formal loss functions. We will encounter numerous illustrations of this, together with indications of the value of using decision-theoretic machinery to then solve the inference problem.

1.3. Expected Loss, Decision Rules, and Risk

As mentioned in the Introduction, we will be involved with decision making in the presence of uncertainty. Hence the actual incurred loss, $L(\theta, a)$, will never be known with certainty (at the time of decision making). A natural method of proceeding in the face of this uncertainty is to consider the "expected" loss of making a decision, and then choose an "optimal" decision with respect to this expected loss. In this section we consider several standard types of expected loss.

1.3.1. Bayesian Expected Loss

From an intuitive viewpoint, the most natural expected loss to consider is one involving the uncertainty in θ, since θ is all that is unknown at the time of making the decision. We have already mentioned that it is possible to treat θ as a random quantity with a probability distribution, and considering expected loss with respect to this probability distribution is eminently sensible (and will indeed be justified in Chapters 2, 3 and 4).

Definition 1. If $\pi^*(\theta)$ is the believed probability distribution of θ at the time of decision making, the *Bayesian expected loss* of an action a is

$$\rho(\pi^*, a) = E^{\pi^*} L(\theta, a) = \int_\Theta L(\theta, a) dF^{\pi^*}(\theta).$$

EXAMPLE 1 (continued). Assume *no* data is obtained, so that the believed distribution of θ_2 is simply $\pi(\theta_2) = 10 I_{(0.1,0.2)}(\theta_2)$. Then

$$\begin{aligned}
\rho(\pi, a) &= \int_0^1 L(\theta_2, a)\pi(\theta_2) d\theta_2 \\
&= \int_0^a 2(a-\theta_2) 10 I_{(0.1,0.2)}(\theta_2) d\theta_2 + \int_a^1 (\theta_2 - a) 10 I_{(0.1,0.2)}(\theta_2) d\theta_2 \\
&= \begin{cases} 0.15 - a & \text{if } a \le 0.1, \\ 15a^2 - 4a + 0.3 & \text{if } 0.1 \le a \le 0.2, \\ 2a - 0.3 & \text{if } a \ge 0.2. \end{cases}
\end{aligned}$$

EXAMPLE 3 (continued). Here

$$\begin{aligned}
\rho(\pi, a_1) &= E^\pi L(\theta, a_1) \\
&= L(\theta_1, a_1)\pi(\theta_1) + L(\theta_2, a_1)\pi(\theta_2) \\
&= (-500)(0.9) + (1000)(0.1) = -350, \\
\rho(\pi, a_2) &= E^\pi L(\theta, a_2) \\
&= L(\theta_1, a_2)\pi(\theta_1) + L(\theta_2, a_2)\pi(\theta_2) \\
&= -300.
\end{aligned}$$

We use π^* in Definition 1, rather than π, because π will usually refer to the *initial* prior distribution for θ, while π^* will typically be the final (*posterior*) distribution of θ after seeing the data (see Chapter 4). Note that it is being implicitly assumed here (and throughout the book) that choice of a will not affect the distribution of θ. When the action does have an effect, one can replace $\pi^*(\theta)$ by $\pi_a^*(\theta)$, and still consider expected loss. See Jeffrey (1983) for development.

1.3.2. Frequentist Risk

The non-Bayesian school of decision theory, which will henceforth be called the *frequentist* or *classical* school, adopts a quite different expected loss based on an average over the random X. As a first step in defining this expected loss, it is necessary to define a decision rule (or decision procedure).

Definition 2. A (nonrandomized) *decision rule* $\delta(x)$ is a function from $\mathscr{X}$ into $\mathscr{A}$. (We will always assume that functions introduced are appropriately "measurable.") If $X = x$ is the observed value of the sample information, then $\delta(x)$ is the action that will be taken. (For a no-data problem, a decision rule is simply an action.) Two decision rules, δ_1 and δ_2, are considered equivalent if $P_\theta(\delta_1(X) = \delta_2(X)) = 1$ for all θ.

EXAMPLE 1 (continued). For the situation of Example 1, $\delta(x) = x/n$ is the standard decision rule for estimating θ_2. (In estimation problems, a decision rule will be called an *estimator*.) This estimator does not make use of the loss function or prior information given in Example 1. It will be seen later how to develop estimators which do so.

EXAMPLE 2 (continued). The decision rule

$$\delta(x) = \begin{cases} a_1 & \text{if } x/n \le 0.05, \\ a_2 & \text{if } x/n > 0.05, \end{cases}$$

is a standard type of rule for this problem.

The frequentist decision-theorist seeks to evaluate, for each θ, how much he would "expect" to lose if he used $\delta(X)$ repeatedly with varying X in the problem. (See Subsection 1.6.2 for justification of this approach.)

Definition 3. The *risk function* of a decision rule $\delta(x)$ is defined by

$$R(\theta, \delta) = E_\theta^X[L(\theta, \delta(X))] = \int_{\mathscr{X}} L(\theta, \delta(x)) dF^X(x|\theta).$$

(For a no-data problem, $R(\theta, \delta) \equiv L(\theta, \delta)$.)

To a frequentist, it is desirable to use a decision rule δ which has small $R(\theta, \delta)$. However, whereas the Bayesian expected loss of an action was a single number, the risk is a *function* on Θ, and since θ is unknown we have a problem in saying what "small" means. The following partial ordering of decision rules is a first step in defining a "good" decision rule.

Definition 4. A decision rule δ_1 is *R-better* than a decision rule δ_2 if $R(\theta, \delta_1) \leq R(\theta, \delta_2)$ for all $\theta \in \Theta$, with strict inequality for some θ. A rule δ_1 is *R-equivalent* to δ_2 if $R(\theta, \delta_1) = R(\theta, \delta_2)$ for all θ.

Definition 5. A decision rule δ is *admissible* if there exists no R-better decision rule. A decision rule δ is *inadmissible* if there does exist an R-better decision rule.

It is fairly clear that an inadmissible decision rule should not be used, since a decision rule with smaller risk can be found. (One might take exception to this statement if the inadmissible decision rule is simple and easy to use, while the improved rule is very complicated and offers only a slight improvement. Another more philosophical objection to this exclusion of inadmissible rules will be presented in Section 4.8.) Unfortunately, there is usually a large class of admissible decision rules for a particular problem. These rules will have risk functions which cross, i.e., which are better in different places. An example of these ideas is given below.

EXAMPLE 4. Assume X is $\mathcal{N}(\theta, 1)$, and that it is desired to estimate θ under loss $L(\theta, a) = (\theta - a)^2$. (This loss is called *squared-error* loss.) Consider the decision rules $\delta_c(x) = cx$. Clearly

$$\begin{aligned} R(\theta, \delta_c) &= E_\theta^X L(\theta, \delta_c(X)) = E_\theta^X (\theta - cX)^2 \\ &= E_\theta^X (c[\theta - X] + [1-c]\theta)^2 \\ &= c^2 E_\theta^X [\theta - X]^2 + 2c(1-c)\theta E_\theta^X [\theta - X] + (1-c)^2\theta^2 \\ &= c^2 + (1-c)^2\theta^2. \end{aligned}$$

Since for $c > 1$,

$$R(\theta, \delta_1) = 1 < c^2 + (1-c)^2\theta^2 = R(\theta, \delta_c),$$

δ_1 is R-better than δ_c for $c > 1$. Hence the rules δ_c are inadmissible for $c > 1$. On the other hand, for $0 \leq c \leq 1$ the rules are noncomparable. For example, the risk functions of the rules δ_1 and $\delta_{1/2}$ are graphed in Figure 1.1. The risk functions clearly cross. Indeed it will be seen later that for $0 \leq c \leq 1$, δ_c is admissible. Thus the "standard" estimator δ_1 is admissible. So, however, is the rather silly estimator δ_0, which estimates θ to be zero no matter what x is observed. (This indicates that while admissibility may be a desirable property for a decision rule, it gives no assurance that the decision rule is reasonable.)

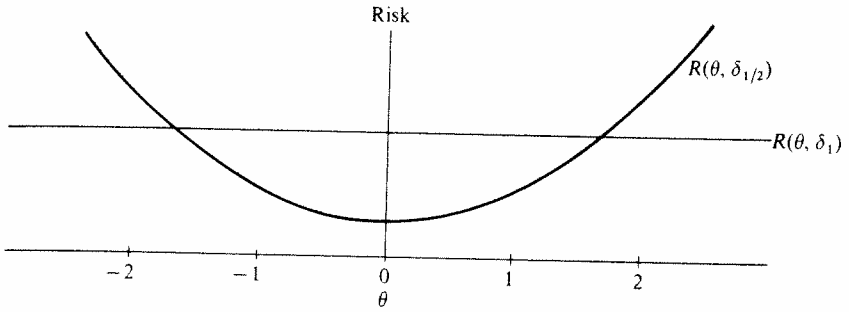

Figure 1.1

EXAMPLE 5. The following is the loss matrix of a particular no-data problem.

	a_1	a_2	a_3
θ_1	1	3	4
θ_2	−1	5	5
θ_3	0	−1	−1

The rule (action) a_2 is R-better than a_3 since $L(\theta_i, a_2) \leq L(\theta_i, a_3)$ for all θ_i, with strict inequality for θ_1. (Recall that, in a no-data problem, the risk is simply the loss.) Hence a_3 is inadmissible. The actions a_1 and a_2 are noncomparable, in that $L(\theta_i, a_1) < L(\theta_i, a_2)$ for θ_1 and θ_2, while the reverse inequality holds for θ_3. Thus a_1 and a_2 are admissible.

In this book we will only consider decision rules with finite risk. More formally, we will assume that the only (nonrandomized) decision rules under consideration are those in the class

$$\mathscr{D} = \{\text{all decision rules } \delta\colon R(\theta, \delta) < \infty \text{ for all } \theta \in \Theta\}.$$

(There are actually technical reasons for allowing infinite risk decision rules in certain abstract settings, but we will encounter no such situations in this book; our life will be made somewhat simpler by not having to worry about infinite risks.)

We defer discussion of the differences between using Bayesian expected loss and the risk function until Section 1.6 (and elsewhere in the book). There is, however, one other relevant expected loss to consider, and that is the expected loss which averages over *both* θ and X.

Definition 6. The *Bayes risk* of a decision rule δ, with respect to a prior distribution π on Θ, is defined as

$$r(\pi, \delta) = E^{\pi}[R(\theta, \delta)].$$

EXAMPLE 4 (continued). Suppose that $\pi(\theta)$ is a $\mathcal{N}(0, \tau^2)$ density. Then, for the decision rule δ_c,

$$r(\pi, \delta_c) = E^{\pi}[R(\theta, \delta_c)] = E^{\pi}[c^2 + (1-c)^2\theta^2]$$
$$= c^2 + (1-c)^2 E^{\pi}[\theta^2] = c^2 + (1-c)^2\tau^2.$$

The Bayes risk of a decision rule will be seen to play an important role in virtually any approach to decision theory.

1.4. Randomized Decision Rules

In some decision situations it is necessary to take actions in a random manner. Such situations most commonly arise when an intelligent adversary is involved. As an example, consider the following game called "matching pennies."

EXAMPLE 6 (Matching Pennies). You and your opponent are to simultaneously uncover a penny. If the two coins match (i.e., are both heads or both tails) you win \$1 from your opponent. If the coins don't match, your opponent wins \$1 from you. The actions which are available to you are a_1—choose heads, or a_2—choose tails. The possible states of nature are θ_1—the opponent's coin is a head, and θ_2—the opponent's coin is a tail. The loss matrix in this game is

	a_1	a_2
θ_1	-1	1
θ_2	1	-1

Both a_1 and a_2 are admissible actions. However, if the game is to be played a number of times, then it would clearly be a very poor idea to decide to use a_1 exclusively or a_2 exclusively. Your opponent would very quickly realize your strategy, and simply choose his action to guarantee victory. Likewise, any patterned choice of a_1 and a_2 could be discerned by an intelligent opponent, who could then develop a winning strategy. The only certain way of preventing ultimate defeat, therefore, is to choose a_1 and a_2 by some random mechanism. A natural way to do this is simply to choose a_1 and a_2 with probabilities p and $1-p$ respectively. The formal definition of such a randomized decision rule follows.

Definition 7. A *randomized decision rule* $\delta^*(x, \cdot)$ is, for each x, a probability distribution on $\mathcal{A}$, with the interpretation that if x is observed, $\delta^*(x, A)$ is

1.5. Decision Principles

In this section we briefly introduce the major methods of actually makin a decision or choosing a decision rule.

1.5.1. The Conditional Bayes Decision Principle

The word "conditional" in the title is explained in Section 1.6, and distinguishes Bayesian analysis using Bayesian expected loss from that using Bayes risk (see also Subsection 4.4.1). It is typically very easy to choose an optimal action when one can determine the Bayesian expected loss, $\rho(\pi^*, a)$, for each a. (Recall that π^* is the believed probability distribution for θ at the time of decision making.)

The Conditional Bayes Principle. *Choose an action $a \in \mathcal{A}$ which minimizes $\rho(\pi^*, a)$ (assuming the minimum is attained). Such an action will be called a* Bayes action *and will be denoted a^{π^*}.*

EXAMPLE 1 (continued). In Subsection 1.3.1 it was shown that

$$\rho(\pi, a) = \begin{cases} 0.15 - a & \text{if } a \le 0.1, \\ 15a^2 - 4a + 0.3 & \text{if } 0.1 \le a \le 0.2, \\ 2a - 0.3 & \text{if } a \ge 0.2. \end{cases}$$

Calculus shows that the minimum of this function over a is $\frac{1}{30}$, and is achieved at $a^{\pi} = \frac{2}{15}$. This, then, would be the estimate for θ_2, the market share of the new drug (assuming no data was available).

EXAMPLE 3 (continued). In Subsection 1.3.1 it was shown that $\rho(\pi, a_1) = -350$ and $\rho(\pi, a_2) = -300$. Clearly a_1 has smaller Bayesian expected loss, and is hence the Bayes action. Thus the risky bonds should be purchased (according to the conditional Bayes principle).

1.5.2. Frequentist Decision Principles

It was remarked in Subsection 1.3.2 that use of risk functions to select a decision rule is difficult, because there are typically many admissible decision rules (i.e., decision rules which can not be dominated in terms of risk). An additional principle must be introduced in order to select a specific rule for use. In classical statistics there are a number of such principles for developing statistical procedures: the maximum likelihood, unbiasedness, minimum variance, and least squares principles to name a few. In decision theory there are also several possible principles that can be used; the three

most important being the Bayes risk principle, the minimax principle, and the invariance principle. In this section the basic goal of each of these three principles is stated. In later chapters the methods of implementation of these principles will be discussed.

I. *The Bayes Risk Principle*

In Subsection 1.3.2 it was seen that an alternative method of involving a prior distribution, π, was to look at $r(\pi, \delta) = E^{\pi}R(\theta, \delta)$, the Bayes risk of δ. Since this is a *number*, we can simply seek a decision rule which minimizes it.

The Bayes Risk Principle. *A decision rule δ_1 is preferred to a rule δ_2 if*

$$r(\pi, \delta_1) < r(\pi, \delta_2).$$

A decision rule which minimizes $r(\pi, \delta)$ is optimal; it is called a Bayes rule, *and will be denoted δ^{π}. The quantity $r(\pi) = r(\pi, \delta^{\pi})$ is then called the* Bayes risk for π.

EXAMPLE 4 (continued). In Subsection 1.3.2 it was calculated that $r(\pi, \delta_c) = c^2 + (1-c)^2\tau^2$, when π is $\mathcal{N}(0, \tau^2)$. Minimizing with respect to c (by differentiating and setting equal to zero) shows that $c_0 = \tau^2/(1+\tau^2)$ is the best value. Thus δ_{c_0} has the smallest Bayes risk among all estimators of the form δ_c. It will be shown in Chapter 4 that δ_{c_0} actually has the smallest Bayes risk among *all* estimators (for the given π). Hence δ_{c_0} is the Bayes rule (or Bayes estimator), and

$$\begin{aligned} r(\pi) = r(\pi, \delta_{c_0}) &= c_0^2 + (1-c_0)^2\tau^2 \\ &= \left(\frac{\tau^2}{1+\tau^2}\right)^2 + \left(\frac{1}{1+\tau^2}\right)^2 \tau^2 = \frac{\tau^2}{1+\tau^2} \end{aligned}$$

is the Bayes risk of π.

EXAMPLE 3 (continued). Since this is a no-data problem, decision rules are simply actions, and the risk function is simply the loss function. Hence the Bayes risk is simply the Bayesian expected loss, and we solved the problem in Subsection 1.5.1.

The above example points out that, in a no-data problem, the Bayes risk principle will give the same answer as the conditional Bayes decision principle. It will, in fact, be seen in Subsection 4.4.1 that this correspondence always holds. (The π^* used in the conditional Bayes decision principle will usually be a data-modified version of the original prior distribution π used in the Bayes risk principle, but the two approaches will yield the same decision.)

II. *The Minimax Principle*

Complete analysis of problems using the minimax principle generally calls for consideration of randomized decision rules. Thus let $\delta^* \in \mathcal{D}^*$ be a randomized rule, and consider the quantity

$$\sup_{\theta \in \Theta} R(\theta, \delta^*).$$

This represents the worst that can happen if the rule δ^* is used. If it is desired to protect against the worst possible state of nature, one is led to using

The Minimax Principle. *A decision rule δ_1^* is preferred to a rule δ_2^* if*

$$\sup_{\theta} R(\theta, \delta_1^*) < \sup_{\theta} R(\theta, \delta_2^*).$$

Definition 10. A rule δ^{*M} is a *minimax decision rule* if it minimizes $\sup_\theta R(\theta, \delta^*)$ among all randomized rules in $\mathcal{D}^*$, i.e., if

$$\sup_{\theta \in \Theta} R(\theta, \delta^{*M}) = \inf_{\delta^* \in \mathcal{D}^*} \sup_{\theta \in \Theta} R(\theta, \delta^*).$$

The quantity on the right-hand side of the above expression is called the *minimax value* of the problem. (Replacing "inf" by "min" and "sup" by "max" shows the origin of the name "minimax.") For no-data problems, the minimax decision rule will be called simply the *minimax action.*

Sometimes it is of interest to determine the best nonrandomized rule according to the minimax principle. If such a best rule exists, it will be called the *minimax nonrandomized rule* (or *minimax nonrandomized action* in no-data problems).

EXAMPLE 4 (continued). For the decision rules δ_c,

$$\sup_{\theta} R(\theta, \delta_c) = \sup_{\theta} [c^2 + (1-c)^2\theta^2] = \begin{cases} 1 & \text{if } c = 1, \\ \infty & \text{if } c \neq 1. \end{cases}$$

Hence δ_1 is best among the rules δ_c, according to the minimax principle. Indeed it will be shown in Chapter 5 that δ_1 is a minimax rule and that 1 is the minimax value for the problem. By using the rule $\delta_1(x) = x$, one can thus ensure that the risk is no worse than 1 (and actually equal to 1 for all θ). Note that the minimax rule and the Bayes rule found earlier are different.

EXAMPLE 3 (continued). Clearly

$$\sup_{\theta} L(\theta, a_1) = \max\{-500, 1000\} = 1000,$$

$$\sup_{\theta} L(\theta, a_2) = \max\{-300, -300\} = -300.$$

Thus a_2 is the minimax nonrandomized action.

EXAMPLE 6 (continued). The randomized rules can be written

$$\delta_p^* = p\langle a_1\rangle + (1-p)\langle a_2\rangle,$$

which recall means that a_1 is to be selected with probability p, and a_2 is to be chosen with probability $1-p$. The loss (and hence risk) of such a rule was shown to be

$$\begin{aligned} R(\theta, \delta_p^*) = L(\theta, \delta_p^*) &= pL(\theta, a_1) + (1-p)L(\theta, a_2) \\ &= \begin{cases} 1-2p & \text{if } \theta = \theta_1, \\ 2p-1 & \text{if } \theta = \theta_2. \end{cases} \end{aligned}$$

Hence

$$\sup_\theta R(\theta, \delta_p^*) = \max\{1-2p, 2p-1\}.$$

Graphing the functions $1-2p$ and $2p-1$ (for $0 \le p \le 1$) and noting that the maximum is always the higher of the two lines, it becomes clear that the minimum value of $\max\{1-2p, 2p-1\}$ is 0, occurring at $p = \frac{1}{2}$. Thus $\delta_{1/2}^*$ is the minimax action and 0 is the minimax value for the problem.

The above example describes a useful technique for solving two action no-data problems. In such situations, the randomized rules are always of the form δ_p^*, and $R(\theta, \delta_p^*)$ can be graphed as a function of p (for $0 \le p \le 1$). If Θ is finite, one need only graph the lines $R(\theta, \delta_p^*)$ for each θ, and note that the highest line segments of the graph form $\sup_\theta R(\theta, \delta_p^*)$. The minimizing value of p can then be seen directly. Techniques for finding minimax rules in more difficult situations will be presented in Chapter 5.

III. *The Invariance Principle*

The invariance principle basically states that if two problems have identical formal structures (i.e., have the same sample space, parameter space, densities, and loss function), then the same decision rule should be used in each problem. This principle is employed, for a given problem, by considering transformations of the problem (say, changes of scale in the unit of measurement) which result in transformed problems of identical structure. The proscription that the decision rules in the original and transformed problem be the same leads to a restriction to so-called "invariant" decision rules. This class of rules will often be small enough so that a "best invariant" decision rule will exist. Chapter 6 will be devoted to the discussion and application of this principle.

The δ which minimizes this will clearly also minimize

$$b(E^{\pi}E_{\theta}^{X}[L(\theta, \delta(X))]) + c = E^{\pi}E_{\theta}^{X}[bL(\theta, \delta(X)) + c],$$

which is the Bayes risk for the linearly transformed loss.

Loss functions are usually bounded from below, i.e.,

$$\inf_{\theta} \inf_{a} L(\theta, a) = B > -\infty.$$

In such situations, the transformed loss function

$$L^*(\theta, a) = L(\theta, a) - B$$

will always be nonnegative. It is convenient to talk in terms of nonnegative losses, so it will often be assumed that the above transformation has been made. Note also that, when derived from a utility function, the loss L^* can be written as

$$L^*(\theta, a) = \sup_{\theta} \sup_{a} U(\theta, a) - U(\theta, a).$$

This actually seems somewhat more sensible as a measure of loss than does the earlier definition in (2.3), in that $L^*(\theta, a)$ measures the true amount "lost" by not having the most favorable possibility occur. The analysis with L^* will not differ from the analysis with L, however. (It could be argued that the true amount "lost" is $\bar{L}(\theta, a) = \sup_a U(\theta, a) - U(\theta, a)$, since one has no control over which θ occurs. This loss is called *regret loss*, and will be discussed in Subsection 5.5.5. Interestingly enough, $\bar{L}$ is equivalent to L and L^* for Bayesian analysis, but can lead to different, and generally more reasonable, results for other decision principles.)

2.4.2. Certain Standard Loss Functions

In making a decision or evaluating a decision rule, the loss function should, ideally, be developed as above. Often, however, analyses of decision rules are carried out for certain "standard" losses. Three of these will be briefly discussed.

I. *Squared-Error Loss*

The loss function $L(\theta, a) = (\theta - a)^2$ is called *squared-error loss*. There are a number of reasons why it is often considered in evaluating decision rules. It was originally used in estimation problems when unbiased estimators of θ were being considered, since $R(\theta, \delta) = E_{\theta}L(\theta, \delta(X)) = E_{\theta}[\theta - \delta(X)]^2$ would then be the variance of the estimator. A second reason for the popularity of squared-error loss is due to its relationship to classical least squares theory. The similarity between the two makes squared-error loss

seem familiar to statisticians. Finally, for most decision analyses, the use of squared-error loss makes the calculations relatively straightforward and simple.

The above justifications for squared-error loss really have very little merit. The question is—does squared-error loss typically reflect the true loss function in a given situation? The initial reaction is, probably, no. As in the discussion of utility theory, one can reason that the loss function should usually be bounded and (at least for large errors) concave. Squared-error loss is neither of these. The convexity of squared-error loss is particularly disturbing. (Large errors are penalized, perhaps, much too severely.)

There are a number of situations, however, in which squared-error loss may be appropriate. For example, in many statistical problems for which a loss symmetric in $(\theta - a)$ is suitable, the exact functional form of the loss is not crucial to the conclusion. Squared-error loss may then be a useful approximation to the true loss. Several problems of this nature will be encountered in later chapters.

Another situation in which squared-error loss can arise is when

$$L(\theta, a) = -U(\theta, a) = -E_{\theta,a}[U(Z)], \tag{2.4}$$

as mentioned earlier. Exercise 14 deals with such a situation, and shows how a loss similar to squared-error loss can occur naturally.

If, more generally, the sample information is quite accurate in a problem with a loss as in (2.4), then $L(\theta, a)$ can frequently be approximated by squared-error loss. For example, assume $Z = h(\theta - a, Y)$, where the distribution of Y does not depend on θ or a. (The reward is thus a function of the accuracy of estimating θ (as measured by $\theta - a$) and some random variable Y which does not depend on θ or a. For example, Y could be a random variable reflecting the future state of the economy.) For convenience, define $g(\theta - a, Y) = U(h(\theta - a, Y))$. Since the sample information about θ is quite accurate, $\theta - a$ will be small. Thus $g(\theta - a, Y)$ can be expanded in a Taylor series about 0, giving

$$g(\theta - a, Y) \cong g(0, Y) + (\theta - a)g'(0, Y) + \tfrac{1}{2}(\theta - a)^2 g''(0, Y).$$

(The derivatives are, of course, with respect to the first argument of g. Higher-order terms will tend to be negligible because they involve higher powers of the small $(\theta - a)$.) Letting

$$K_1 = -E^Y[g(0, Y)], \qquad K_2 = -E[g'(0, Y)], \quad \text{and} \quad K_3 = -\tfrac{1}{2}E[g''(0, Y)],$$

it follows that

$$L(\theta, a) = -E[U(Z)] \cong K_1 + K_2(\theta - a) + K_3(\theta - a)^2.$$

Completing squares gives

$$L(\theta, a) \cong K_3\left(\theta - a + \frac{K_2}{2K_3}\right)^2 + \left(K_1 - \frac{K_2^2}{4K_3}\right).$$

Providing K_3 is a positive constant, this loss is equivalent (for decisic making) to the transformed loss

$$L(\theta, a) = \left(\theta - a + \frac{K_2}{2K_3}\right)^2.$$

This would be squared-error loss if it weren't for the constant $K_2/2K$ When $K_2 = 0$ (which would occur if $g(0, y)$ was a symmetric function of and Y had a symmetric distribution), there is no problem. Otherwis however, the constant represents the fact that either overestimation c underestimation (depending on the sign of K_2) is desired.

To analyze a decision problem with

$$L(\theta, a) = (\theta - a + c)^2,$$

merely consider the new action space

$$\mathscr{A}^* = \{a - c: a \in \mathscr{A}\}.$$

For $a^* \in \mathscr{A}^*$, the loss corresponding to L is $L^*(\theta, a^*) = (\theta - a^*)^2$. Th analysis in this transformed problem is thus done with squared-error loss If δ^* is an optimal decision rule in the transformed problem, then $\delta = \delta^* +$ will be optimal in the original problem.

A generalization of squared-error loss, which is of interest, is

$$L(\theta, a) = w(\theta)(\theta - a)^2.$$

This loss is called *weighted squared-error loss*, and has the attractive feature of allowing the squared error, $(\theta - a)^2$, to be weighted by a function of θ This will reflect the fact that a given error in estimation often varies in harm according to what θ happens to be.

The final variant of squared-error loss which will be considered is quadratic loss. If $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ is a vector to be estimated by $\mathbf{a} = (a_1, \ldots, a_p)'$ and $\mathbf{Q}$ is a $p \times p$ positive definite matrix, then

$$L(\boldsymbol{\theta}, \mathbf{a}) = (\boldsymbol{\theta} - \mathbf{a})'\mathbf{Q}(\boldsymbol{\theta} - \mathbf{a})$$

is called *quadratic loss.* When $\mathbf{Q}$ is diagonal, this reduces to

$$L(\boldsymbol{\theta}, \mathbf{a}) = \sum_{i=1}^{p} q_i(\theta_i - a_i)^2,$$

and is a natural extension of squared-error loss to the multivariate situation.

II. *Linear Loss*

When the utility function is approximately linear (as is often the case over a reasonable segment of the reward space), the loss function will tend to

be linear. Thus of interest is the *linear loss*

$$L(\theta, a) = \begin{cases} K_0(\theta - a) & \text{if } \theta - a \geq 0, \\ K_1(a - \theta) & \text{if } \theta - a < 0. \end{cases}$$

The constants K_0 and K_1 can be chosen to reflect the relative importance of underestimation and overestimation. These constants will usually be different. When they are equal, the loss is equivalent to

$$L(\theta, a) = |\theta - a|,$$

which is called *absolute error loss.* If K_0 and K_1 are functions of θ, the loss will be called *weighted linear loss.* Linear loss (or weighted linear loss) is quite often a useful approximation to the true loss.

III. "0-1" *Loss*

In the two-action decision problem (of which hypothesis testing is an example) it is typically the case that a_0 is "correct" if $\theta \in \Theta_0$, and a_1 is correct if $\theta \in \Theta_1$. (This could correspond to testing H_0: $\theta \in \Theta_0$ versus H_1: $\theta \in \Theta_1$.) The loss

$$L(\theta, a_i) = \begin{cases} 0 & \text{if } \theta \in \Theta_i, \\ 1 & \text{if } \theta \in \Theta_j \qquad (j \neq i), \end{cases}$$

is called "0-1" *loss.* In words, this loss is zero if a correct decision is made, and 1 if an incorrect decision is made. The interest in this loss arises from the fact that, in a testing situation, the risk function of a decision rule (or test) $\delta(x)$ is simply

$$R(\theta, \delta) = E_\theta[L(\theta, \delta(X))] = P_\theta(\delta(X) \text{ is the incorrect decision}).$$

This is either a probability of Type I or Type II error, depending on whether $\theta \in \Theta_0$ or $\theta \in \Theta_1$. And, similarly from the conditional perspective, the Bayesian expected loss is

$$\rho(\pi^*, a_i) = \int L(\theta, a_i) dF^{\pi^*}(\theta) = 1 - P^{\pi^*}(\theta \in \Theta_i),$$

which is one minus the actual (subjective) *probability* that H_i is true (in light of the probability distribution, π^*, for θ).

In practice, "0-1" loss will rarely be a good approximation to the true loss. More realistic losses are

$$L(\theta, a_i) = \begin{cases} 0 & \text{if } \theta \in \Theta_i, \\ k_i & \text{if } \theta \in \Theta_j \qquad (i \neq j), \end{cases}$$

and

$$L(\theta, a_i) = \begin{cases} 0 & \text{if } \theta \in \Theta_i, \\ k_i(\theta) & \text{if } \theta \in \Theta_j \qquad (i \neq j). \end{cases}$$

This last type of loss, with $k_i(\theta)$ being an increasing function of tl "distance" of the true θ from Θ_i, is particularly reasonable, in that the har suffered by an incorrect decision will usually depend on the severity of tl mistake. Actually, even when a "correct" decision is made, $L(\theta, a)$ m very well be nonzero, so that the full generality of the loss may be neede Note, however, that only two functions, $L(\theta, a_0)$ and $L(\theta, a_1)$, must l determined.

For the most part, examples in the book will use the above three loss or variants of them. This is done mainly to make the calculations relative easy. It should be emphasized that these losses need not necessarily l suitable for a given problem. Indeed, about the only way to decide if th are reasonable is to do a utility analysis.

2.4.3. For Inference Problems

In Section 1.2 we discussed, in general terms, the question of the applicabili of decision theory to "statistical inference." We will not pursue the issu of whether or not inference problems really exist (i.e., the argument th *any* conclusion will ultimately be used for something and one should t to anticipate such uses), and will instead concentrate on how decision theo can aid in the process of choosing an inference in a more formal sense.

The most common application of decision theory to inference problen is through the representation of common inference measures as a risk Bayesian expected loss. For instance, in Subsection 2.4.2 it was shown th frequentist error probabilities in testing correspond to the risk function f "0-1" loss. Another similar example is the confidence set scenario: if denotes a confidence rule (when x is observed, the set $C(x) \subset \Theta$ will b presented as the confidence set for θ), and one considers the loss functic

$$L(\theta, C(x)) = 1 - I_{C(x)}(\theta) = \begin{cases} 1 & \text{if } \theta \notin C(x), \\ 0 & \text{if } \theta \in C(x), \end{cases}$$

then

$$R(\theta, C) = E_\theta[1 - I_{C(X)}(\theta)] = 1 - P_\theta(C(X) \text{ contains } \theta),$$

which is one minus the frequentist coverage probability. Likewise, for distribution π^* for θ,

$$\rho(\pi^*, C(x)) = E^{\pi^*}[1 - I_{C(x)}(\theta)] = 1 - P^{\pi^*}(\theta \in C(x)),$$

which is one minus the actual (subjective) probability that θ is in the specifi set $C(x)$ (for the observed x).

A number of other frequentist and Bayesian inference measures can als be given a formal interpretation as a risk or a Bayesian expected loss. Th task of finding a good inference procedure (or conditional inference) the

values of θ_1, and also determine $\pi_1(\theta_1)$ (the marginal density of θ_1). Since $\pi(\theta_1, \theta_2) = \pi_1(\theta_1)\pi(\theta_2|\theta_1)$, the joint prior density can thus be approximated.

There has been extensive study of the ability of people (experts and nonexperts) to elicit probability distributions. The studies show that untrained (or unpracticed) elicitors do quite poorly, primarily because of substantial overconfidence concerning their prior knowledge (i.e., the elicited distributions are much too tightly concentrated). Discussion and many references can be found in Alpert and Raiffa (1982) and in Kahneman, Slovic, and Tversky (1982). Besides the obvious solution of attempting to better train people in probability elicitation, one can attempt to partially alleviate the problem through study of "Bayesian robustness" (see Section 4.7).

3.3. Noninformative Priors

3.3.1. Introduction

Because of the compelling reasons to perform a conditional analysis and the attractiveness of using Bayesian machinery to do so (see Chapter 4), there have been attempts to use the Bayesian approach even when no (or minimal) prior information is available. What is needed in such situations is a *noninformative prior*, by which is meant a prior which contains no information about θ (or more crudely which "favors" no possible values of θ over others). For example, in testing between two simple hypotheses, the prior which gives probability $\frac{1}{2}$ to each of the hypotheses is clearly noninformative. The following is a more complex example.

EXAMPLE 4. Suppose the parameter of interest is a normal mean θ, so that the parameter space is $\Theta = (-\infty, \infty)$. If a noninformative prior density is desired, it seems reasonable to give equal weight to all possible values of θ. Unfortunately, if $\pi(\theta) = c > 0$ is chosen, then π has infinite mass (i.e. $\int \pi(\theta)d\theta = \infty$) and is not a proper density. Nevertheless, such π can be successfully worked with. The choice of c is unimportant, so that typically the noninformative prior density for this problem is chosen to be $\pi(\theta) = 1$. This is often called the *uniform density on* R^1, and was introduced and used by Laplace (1812).

As in the above example, it will frequently happen that the natural noninformative prior is an *improper prior*, namely one which has infinite mass. Let us now consider the problem of determining noninformative priors.

The simplest situation to consider is when Θ is a finite set, consisting of say n elements. The obvious noninformative prior is to then give each

element of Θ probability $1/n$. One might generalize this (and Example 4) to infinite Θ by giving each $\theta \in \Theta$ equal density, arriving at the uniform noninformative prior $\pi(\theta) \equiv c$. Although this was routinely done by Laplace (1812), it came under severe (though unjustified) criticism because of a lack of invariance under transformation (see Jaynes (1983) for discussion).

EXAMPLE 4 (continued). Instead of considering θ, suppose the problem had been parameterized in terms of $\eta = \exp\{\theta\}$. This is a one-to-one transformation, and so should have no bearing on the ultimate answer. But, if $\pi(\theta)$ is the density for θ, then the corresponding density for η is (noting that $d\theta/d\eta = d \log \eta/d\eta = \eta^{-1}$ gives the Jacobian)

$$\pi^*(\eta) = \eta^{-1}\pi(\log \eta).$$

Hence, if the noninformative prior for θ is chosen to be constant, we should choose the noninformative prior for η to be proportional to η^{-1} to maintain consistency (and arrive at the same answers in either parameterization). Thus we cannot maintain consistency and choose both the noninformative prior for θ and that for η to be constant.

It could, perhaps, be argued that one usually chooses the most intuitively reasonable parameterization, and that a lack of prior information should correspond to a constant density in this parameterization, but the argument would be hard to defend in general. The lack of invariance of the constant prior has led to a search for noninformative priors which *are* appropriately invariant under transformations. Before discussing the general case in Subsections 3.3.3 and 3.3.4, we present two important examples in Subsection 3.3.2.

3.3.2. Noninformative Priors for Location and Scale Problems

Efforts to derive noninformative priors through consideration of transformations of a problem had its beginnings with Jeffreys (cf. Jeffreys (1961)). It has been extensively used in Hartigan (1964), Jaynes (1968, 1983), Villegas (1977, 1981, 1984), and elsewhere. We present here two illustrations of the idea.

EXAMPLE 5 (Location Parameters). Suppose that $\mathscr{X}$ and Θ are subsets of R^p, and that the density of $\mathbf{X}$ is of the form $f(\mathbf{x}-\boldsymbol{\theta})$ (i.e., depends only on $(\mathbf{x}-\boldsymbol{\theta})$). The density is then said to be a *location density*, and $\boldsymbol{\theta}$ is called a *location parameter* (or sometimes a *location vector* when $p \geq 2$). The $\mathscr{N}(\theta, \sigma^2)$ (σ^2 fixed), $\mathscr{T}(\alpha, \mu, \sigma^2)$ (α and σ^2 fixed), $\mathscr{C}(\alpha, \beta)$ (β fixed), and $\mathscr{N}_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ ($\boldsymbol{\Sigma}$ fixed) densities are all examples of location densities. Also, a

sample of independent identically distributed random variables is said to be from a location density if their common density is a location density.

To derive a noninformative prior for this situation, imagine that, instead of observing $\mathbf{X}$, we observe the random variable $\mathbf{Y}=\mathbf{X}+\mathbf{c}$ ($\mathbf{c}\in R^p$). Defining $\boldsymbol{\eta}=\boldsymbol{\theta}+\mathbf{c}$, it is clear that $\mathbf{Y}$ has density $f(\mathbf{y}-\boldsymbol{\eta})$. If now $\mathscr{X}=\Theta=R^p$, then the sample space and parameter space for the $(\mathbf{Y},\boldsymbol{\eta})$ problem are also R^p. The $(\mathbf{X},\boldsymbol{\theta})$ and $(\mathbf{Y},\boldsymbol{\eta})$ problems are thus identical in structure, and it seems reasonable to insist that they have the same noninformative prior. (Another way of thinking of this is to note that observing $\mathbf{Y}$ really amounts to observing $\mathbf{X}$ with a different unit of measurement, one in which the "origin" is $\mathbf{c}$ and not zero. Since the choice of an origin for a unit of measurement is quite arbitrary, the noninformative prior should perhaps be independent of this choice.)

Letting π and π^* denote the noninformative priors in the $(\mathbf{X},\boldsymbol{\theta})$ and $(\mathbf{Y},\boldsymbol{\eta})$ problems respectively, the above argument implies that π and π^* should be equal, i.e., that

$$P^{\pi}(\boldsymbol{\theta}\in A)=P^{\pi^*}(\boldsymbol{\eta}\in A) \tag{3.1}$$

for any set A in R^p. Since $\boldsymbol{\eta}=\boldsymbol{\theta}+\mathbf{c}$, it should also be true (by a simple change of variables) that

$$P^{\pi^*}(\boldsymbol{\eta}\in A)=P^{\pi}(\boldsymbol{\theta}+\mathbf{c}\in A)=P^{\pi}(\boldsymbol{\theta}\in A-\mathbf{c}), \tag{3.2}$$

where $A-\mathbf{c}=\{\mathbf{z}-\mathbf{c}\colon \mathbf{z}\in A\}$. Combining (3.1) and (3.2) shows that π should satisfy

$$P^{\pi}(\boldsymbol{\theta}\in A)=P^{\pi}(\boldsymbol{\theta}\in A-\mathbf{c}). \tag{3.3}$$

Furthermore, this argument applies no matter which $\mathbf{c}\in R^p$ is chosen, so that (3.3) should hold for all $\mathbf{c}\in R^p$. Any π satisfying this relationship is said to be a *location invariant* prior.

Assuming that the prior has a density, we can write (3.3) as

$$\int_A \pi(\boldsymbol{\theta})d\boldsymbol{\theta}=\int_{A-c}\pi(\boldsymbol{\theta})d\boldsymbol{\theta}=\int_A \pi(\boldsymbol{\theta}-\mathbf{c})d\boldsymbol{\theta}.$$

If this is to hold for all sets A, it can be shown that it must be true that

$$\pi(\boldsymbol{\theta})=\pi(\boldsymbol{\theta}-\mathbf{c})$$

for all $\boldsymbol{\theta}$. Setting $\boldsymbol{\theta}=\mathbf{c}$ thus gives

$$\pi(\mathbf{c})=\pi(\mathbf{0}).$$

Recall, however, that this should hold for all $\mathbf{c}\in R^p$. The conclusion is that π must be a constant function. It is convenient to choose the constant to be 1, so the *noninformative prior density for a location parameter* is $\pi(\boldsymbol{\theta})=1$. (This conclusion can be shown to follow from (3.3), even without the assumption that the prior has a density.)

EXAMPLE 6 (Scale Parameters). A (one-dimensional) *scale density* is a density of the form

$$\sigma^{-1} f\left(\frac{x}{\sigma}\right),$$

where $\sigma > 0$. The parameter σ is called a *scale parameter.* The $\mathcal{N}(0, \sigma^2)$, $\mathcal{T}(\alpha, 0, \sigma^2)$ (α fixed), and $\mathcal{G}(\alpha, \beta)$ (α fixed) densities are all examples of scale densities. Also, a sample of independent identically distributed random variables is said to be from a scale density if their common density is a scale density.

To derive a noninformative prior for this situation, imagine that, instead of observing X, we observe the random variable $Y = cX$ ($c > 0$). Defining $\eta = c\sigma$, an easy calculation shows that the density of Y is $\eta^{-1} f(y/\eta)$. If now $\mathcal{X} = R^1$ or $\mathcal{X} = (0, \infty)$, then the sample and parameter spaces for the (X, σ) problem are the same as those for the (Y, η) problem. The two problems are thus identical in structure, which again indicates that they should have the same noninformative prior. (Here the transformation can be thought of as simply a change in the scale of measurement, from say inches to feet.) Letting π and π^* denote the priors in the (X, σ) and (Y, η) problems, respectively, this means that the equality

$$P^{\pi}(\sigma \in A) = P^{\pi^*}(\eta \in A)$$

should hold for all $A \subset (0, \infty)$. Since $\eta = c\sigma$, it should also be true that

$$P^{\pi^*}(\eta \in A) = P^{\pi}(\sigma \in c^{-1}A),$$

where $c^{-1}A = \{c^{-1}z\colon z \in A\}$. Putting these together, it follows that π should satisfy

$$P^{\pi}(\sigma \in A) = P^{\pi}(\sigma \in c^{-1}A). \tag{3.4}$$

This should hold for all $c > 0$, and any distribution π for which this is true is called *scale invariant.*

The mathematical analysis of (3.4) proceeds as in the preceding example. Write (3.4) (assuming densities) as

$$\int_A \pi(\sigma)\,d\sigma = \int_{c^{-1}A} \pi(\sigma)\,d\sigma = \int_A \pi(c^{-1}\sigma)c^{-1}\,d\sigma,$$

and conclude that, for this to hold for all A, it must be true that

$$\pi(\sigma) = c^{-1}\pi(c^{-1}\sigma)$$

for all σ. Choosing $\sigma = c$, it follows that

$$\pi(c) = c^{-1}\pi(1).$$

Setting $\pi(1) = 1$ for convenience, and noting that the above equality must hold for all $c > 0$, it follows that a reasonable *noninformative prior for a*

scale parameter is $\pi(\sigma) = \sigma^{-1}$. Observe that this is also an improper prior, since $\int_0^\infty \sigma^{-1}\, d\sigma = \infty$.

An interesting natural application of this noninformative prior (discussed in Rosenkrantz (1977)) is to the "table entry" problem. This problem arose from the study of positive entries in various "natural" numerical tables, such as tables of positive physical constants, tables of population sizes of cities, etc. The problem is to determine the relative frequencies of the integers 1 through 9 in the *first* significant digit of the table entries. Intuitively, one might expect each digit to occur $\frac{1}{9}$ of the time. Instead, it has been found that i $(i = 1, \ldots, 9)$ occurs with a relative frequency of about $\log(1 + i^{-1})/\log 10$.

Numerous explanations of this phenomenon have been proposed. The explanation of interest to us is that, since the scale of measurement of these positive entries is quite arbitrary, one might expect the distribution of table entries to be scale invariant. This suggests using $\pi(\sigma) = \sigma^{-1}$ to describe the distribution of table entries σ. Since this is not a proper density, it cannot be formally considered the distribution of σ. Nevertheless, one can use it (properly normalized) to represent the actual distribution of σ on any interval (a, b) where $0 < a < b < \infty$. For example, consider the interval $(1, 10)$. Then the properly normalized version of π is $\pi(\sigma) = \sigma^{-1}/\log 10$. In this region, σ will have first digit i when it lies in the interval $[i, i+1)$. The probability of this is

$$p_i = \int_i^{i+1} [\sigma \log 10]^{-1}\, d\sigma = \frac{\log(i+1) - \log i}{\log 10} = \frac{\log(1 + i^{-1})}{\log 10},$$

which is precisely the relative frequency that is observed in reality. One might object to the rather arbitrary choice of the interval $(1, 10)$, but the following compelling result can be obtained: if the p_i are calculated for an arbitrary interval (a, b), then as $a \to 0$ or $b \to \infty$ or both, the p_i will converge to the values $\log(1 + i^{-1})/\log 10$. This apparent natural occurrence of the noninformative prior may be mere coincidence, but it is intriguing.

The derivations of the noninformative priors in the two previous examples should not be considered completely compelling. There is indeed a logical flaw in the analyses, caused by the fact that the final priors are improper. The difficulty arises in the argument that if two problems have identical structure, they should have the same noninformative prior. The problem here is that, when improper, noninformative priors are not unique. Multiplying an improper prior π by a constant K results in an equivalent prior, in the sense that all decisions and inferences in Bayesian analysis will be identical for the priors π and $K\pi$. Thus there is no reason to insist that π^* and π, in Examples 5 and 6, must be identical. They need only be constant multiples of each other. In Example 5, for instance, this milder restriction will give, in place of (3.1), the relationship

$$P^{\pi}(A) = h(\mathbf{c}) P^{\pi^*}(A),$$

where $h(\mathbf{c})$ is some positive function and P is to be interpreted more liberally as a "measure." The analogue of Equation (3.2), namely

$$P^{\pi^*}(A) = P^{\pi}(A-\mathbf{c}),$$

should remain valid, being as it merely specifies that a change of variables should not affect the measure. Combining these equations gives, in place of (3.3), the relationship

$$P^{\pi}(A) = h(\mathbf{c})P^{\pi}(A-\mathbf{c}).$$

In integral form this becomes

$$\int_A \pi(\boldsymbol{\theta})d\boldsymbol{\theta} = h(\mathbf{c})\int_{A-\mathbf{c}} \pi(\boldsymbol{\theta})d\boldsymbol{\theta} = h(\mathbf{c})\int_A \pi(\boldsymbol{\theta}-\mathbf{c})d\boldsymbol{\theta}.$$

For this to hold for all A, it must be true that

$$\pi(\boldsymbol{\theta}) = h(\mathbf{c})\pi(\boldsymbol{\theta}-\mathbf{c}).$$

Setting $\boldsymbol{\theta} = \mathbf{c}$, it follows that $h(\mathbf{c}) = \pi(\mathbf{c})/\pi(\boldsymbol{\theta})$. The conclusion is that π need only satisfy the functional equation

$$\pi(\boldsymbol{\theta}-\mathbf{c}) = \frac{\pi(\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\mathbf{c})}. \tag{3.5}$$

There are many improper priors, besides the uniform, which satisfy this relationship. An example is $\pi(\boldsymbol{\theta}) = \exp\{\boldsymbol{\theta}'\mathbf{z}\}$, where $\mathbf{z}$ is any fixed vector. A prior satisfying (3.5) is called *relatively location invariant.*

The above problem will be encountered in virtually any situation for which improper noninformative priors must be considered. There will be a wide class of "logically possible" noninformative priors. (See Hartigan (1964) and Stein (1965) for general results of this nature.) Selecting from among this class can be difficult. For certain statistical problems, a natural choice does exist, namely the right invariant Haar measure. A general discussion of this concept requires group theory, and will be delayed until Chapter 6. It is a fact, however, that the noninformative priors given in Examples 5 and 6 are the right invariant Haar measures.

3.3.3. Noninformative Priors in General Settings

The type of argument given in Subsection 3.3.2 requires a special "group invariant" structure to the problem (see Chapter 6). For more general problems, various (somewhat *ad hoc*) suggestions have been advanced for determining a noninformative prior. The most widely used method is that of Jeffreys (1961), which is to choose

$$\pi(\theta) = [I(\theta)]^{1/2} \tag{3.6}$$

as a noninformative prior, where $I(\theta)$ is the expected Fisher information;

under commonly satisfied assumptions (cf. Lehmann (1983)) this is give

$$I(\theta) = -E_\theta\left[\frac{\partial^2 \log f(X|\theta)}{\partial \theta^2}\right].$$

It is easy to calculate that the noninformative priors in Examples 5 a are indeed proportional to (3.6).

If $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^t$ is a vector, Jeffreys (1961) suggests the use of

$$\pi(\theta) = [\det \mathbf{I}(\boldsymbol{\theta})]^{1/2} \quad (.$$

(here "det" stands for determinant), where $\mathbf{I}(\boldsymbol{\theta})$ is the $(p \times p)$ Fisher infor tion matrix; under commonly satisfied assumptions, this is the matrix v (i, j) element

$$I_{ij}(\boldsymbol{\theta}) = -E_{\boldsymbol{\theta}}\left[\frac{\partial^2}{\partial \theta_i\, \partial \theta_j} \log f(X|\boldsymbol{\theta})\right].$$

EXAMPLE 7 (Location-Scale Parameters). A *location-scale* density i density of the form $\sigma^{-1} f((x-\theta)/\sigma)$, where $\theta \in R^1$ and $\sigma > 0$ are unknown parameters. The $\mathcal{N}(\theta, \sigma^2)$ and the $\mathcal{T}(\alpha, \theta, \sigma^2)$ (α fixed) densi are the crucial examples of location-scale densities. A sample of independ identically distributed random variables is said to be from a location-sc density if their common density is a location-scale density.

Working with the normal distribution for simplicity, and noting t $\boldsymbol{\theta} = (\theta, \sigma)$ in this problem, the Fisher information matrix is

$$\begin{aligned}
\mathbf{I}(\boldsymbol{\theta}) &= -E_{\boldsymbol{\theta}}\begin{pmatrix} \dfrac{\partial^2}{\partial \theta^2}\left(-\log\sigma - \dfrac{(X-\theta)^2}{2\sigma^2}\right) & \dfrac{\partial^2}{\partial \theta\, \partial \sigma}\left(-\log\sigma - \dfrac{(X-\theta)^2}{2\sigma^2}\right) \\ \dfrac{\partial^2}{\partial \theta\, \partial \sigma}\left(-\log\sigma - \dfrac{(X-\theta)^2}{2\sigma^2}\right) & \dfrac{\partial^2}{\partial \sigma^2}\left(-\log\sigma - \dfrac{(X-\theta)^2}{2\sigma^2}\right) \end{pmatrix} \\
&= -E_{\boldsymbol{\theta}}\begin{pmatrix} -1/\sigma^2 & 2(\theta - X)/\sigma^3 \\ 2(\theta - X)/\sigma^3 & 1/\sigma^2 - 3(X-\theta)^2/\sigma^4 \end{pmatrix} \\
&= \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{pmatrix}.
\end{aligned}$$

Hence (3.7) becomes

$$\pi(\boldsymbol{\theta}) = \left(\frac{1}{\sigma^2} \cdot \frac{2}{\sigma^2}\right)^{1/2} \propto \frac{1}{\sigma^2}.$$

(This is improper, so we can again ignore any multiplicative constants.) is interesting that this prior can also be shown to arise from an invarian under-transformation argument, as in Subsection 3.3.2.

A noninformative prior for this situation could, alternatively, be deriv by assuming *independence* of θ and σ, and multiplying the noninformati priors obtained in Examples 5 and 6. The result is

$$\pi(\theta, \sigma) = \frac{1}{\sigma}; \quad (3$$