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 $\Theta = \{\theta_2 \colon 0 \le \theta_2 \le 1\} = [0, 1]$. 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 $\mathscr{A} = [0, 1]$. (Usually $\mathscr{A} = \Theta$ 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 $\mathscr{B}(n, \theta_2)$ (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 $\mathscr{U}(0.1, 0.2)$ 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 $\mathscr{B}(n, \theta)$ 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 $\mathscr{B}e(0.05, 1)$ 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 $\mathscr{A} = \{a_1, a_2\}$, 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 $\mathscr{A}$ 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

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

(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.)

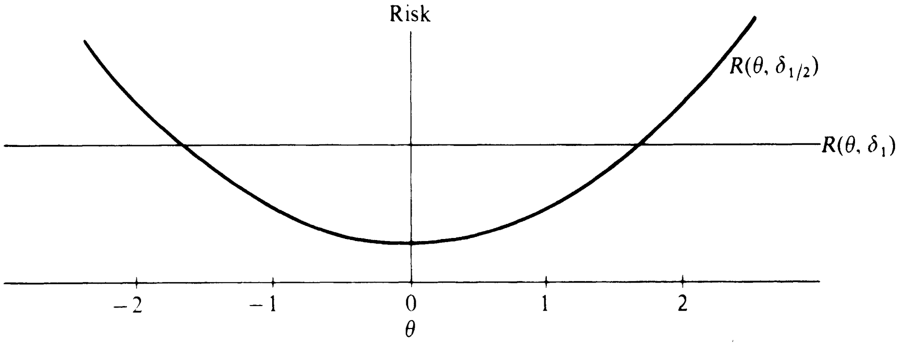

Figure 1.1

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,

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

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 $\mathscr{A}$, with the interpretation that if x is observed, $\delta^*(x, A)$ is