2

Decision theory

In this chapter we give an account of the main ideas of decision theory. Our motivation for beginning our account of statistical inference here is simple. As we have noted, decision theory requires formal specification of all elements of an inference problem, so starting with a discussion of decision theory allows us to set up notation and basic ideas that run through the remainder of the book in a formal but easy manner. In later chapters, we will develop the specific techniques of statistical inference that are central to the three paradigms of inference. In many cases these techniques can be seen as involving the removal of certain elements of the decision theory structure, or focus on particular elements of that structure.

Central to decision theory is the notion of a set of *decision rules* for an inference problem. Comparison of different decision rules is based on examination of the *risk functions* of the rules. The risk function describes the expected *loss* in use of the rule, under hypothetical repetition of the sampling experiment giving rise to the data x, as a function of the *parameter* of interest. Identification of an optimal rule requires introduction of fundamental principles for discrimination between rules, in particular the *minimax* and *Bayes* principles.

2.1 Formulation

A full description of a statistical decision problem involves the following formal elements:

1 A *parameter space* Θ, which will usually be a subset of $\mathbb{R}^d$ for some $d \geq 1$, so that we have a vector of d unknown parameters. This represents the set of possible unknown states of nature. The unknown parameter value $\theta \in \Theta$ is the quantity we wish to make inference about.

2 A *sample space* $\mathcal{X}$, the space in which the data x lie. Typically we have n observations, so the data, a generic element of the sample space, are of the form $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$.

3 A *family of probability distributions* on the sample space $\mathcal{X}$, indexed by values $\theta \in \Theta$, $\{\mathbb{P}_\theta(x),\ x \in \mathcal{X},\ \theta \in \Theta\}$. In nearly all practical cases this will consist of an assumed parametric family $f(x;\theta)$, of probability mass functions for x (in the discrete case), or probability density functions for x (in the continuous case).

4 An *action space* $\mathcal{A}$. This represents the set of all actions or decisions available to the experimenter.

Examples of action spaces include the following:

(a) In a hypothesis testing problem, where it is necessary to decide between two hypotheses H_0 and H_1, there are two possible actions corresponding to 'accept H_0' and

'accept H_1'. So here $\mathcal{A} = \{a_0, a_1\}$, where a_0 represents accepting H_0 and a_1 represents accepting H_1.

(b) In an estimation problem, where we want to estimate the unknown parameter value θ by some function of $x = (x_1, \ldots, x_n)$, such as $\bar{x} = \frac{1}{n}\sum x_i$ or $s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$ or $x_1^3 + 27\sin(\sqrt{x_2})$, etc., we should allow ourselves the possibility of estimating θ by any point in Θ. So, in this context we typically have $\mathcal{A} \equiv \Theta$.

(c) However, the scope of decision theory also includes things such as 'approve Mr Jones' loan application' (if you are a bank manager) or 'raise interest rates by 0.5%' (if you are the Bank of England or the Federal Reserve), since both of these can be thought of as actions whose outcome depends on some unknown state of nature.

5 A *loss function* $L : \Theta \times \mathcal{A} \to \mathbb{R}$ links the action to the unknown parameter. If we take action $a \in \mathcal{A}$ when the true state of nature is $\theta \in \Theta$, then we incur a loss $L(\theta, a)$.

Note that losses can be positive or negative, a negative loss corresponding to a gain. It is a convention that we formulate the theory in terms of trying to minimise our losses rather than trying to maximise our gains, but obviously the two come to the same thing.

6 A set $\mathcal{D}$ of *decision rules*. An element $d : \mathcal{X} \to \mathcal{A}$ of $\mathcal{D}$ is such that each point x in $\mathcal{X}$ is associated with a specific action $d(x) \in \mathcal{A}$.

For example, with hypothesis testing, we might adopt the rule: 'Accept H_0 if $\bar{x} \leq 5.7$, otherwise accept H_1.' This corresponds to a decision rule,

$$d(x) = \begin{cases} a_0 & \text{if } \bar{x} \leq 5.7, \\ a_1 & \text{if } \bar{x} > 5.7. \end{cases}$$

2.2 The risk function

For parameter value $\theta \in \Theta$, the risk associated with a decision rule d based on random data X is defined by

$$\begin{aligned} R(\theta, d) &= \mathbb{E}_\theta L(\theta, d(X)) \\ &= \begin{cases} \int_{\mathcal{X}} L(\theta, d(x)) f(x;\theta)\, dx & \text{for continuous } X, \\ \sum_{x \in \mathcal{X}} L(\theta, d(x)) f(x;\theta) & \text{for discrete } X. \end{cases} \end{aligned}$$

So, we are treating the observed data x as the realised value of a random variable X with density or mass function $f(x;\theta)$, and defining the risk to be the expected loss, the expectation being with respect to the distribution of X for the particular parameter value θ.

The key notion of decision theory is that different decision rules should be compared by comparing their risk functions, as functions of θ. Note that we are explicitly invoking the repeated sampling principle here, the definition of risk involving hypothetical repetitions of the sampling mechanism that generated x, through the assumed distribution of X.

When a loss function represents the real loss in some practical problem (as opposed to some artificial loss function being set up in order to make the statistical decision problem well defined) then it should really be measured in units of 'utility' rather than actual money. For example, the expected return on a UK lottery ticket is less than the £1 cost of the ticket; if everyone played so as to maximise their expected gain, nobody would ever buy a lottery ticket! The reason that people still buy lottery tickets, translated into the language of

statistical decision theory, is that they subjectively evaluate the very small chance of winning, say, £1 000 000 as worth more than a fixed sum of £1, even though the chance of actually winning the £1 000 000 is appreciably less than 1/1 000 000. There is a formal theory, known as utility theory, which asserts that, provided people behave rationally (a considerable assumption in its own right!), then they will always act *as if* they were maximising the expected value of a function known as the utility function. In the lottery example, this implies that we subjectively evaluate the possibility of a massive prize, such as £1 000 000, to be worth more than 1 000 000 times as much as the relatively paltry sum of £1. However in situations involving monetary sums of the same order of magnitude, most people tend to be risk averse. For example, faced with a choice between:

Offer 1: Receive £10 000 with probability 1;

and

Offer 2: Receive £20 000 with probability $\frac{1}{2}$, otherwise receive £0,

most of us would choose Offer 1. This means that, in utility terms, we consider £20 000 as worth less than twice as much as £10 000. Either amount seems like a very large sum of money, and we may not be able to distinguish the two easily in our minds, so that the lack of risk involved in Offer 1 makes it appealing. Of course, if there was a specific reason why we really needed £20 000, for example because this was the cost of a necessary medical operation, we might be more inclined to take the gamble of Offer 2.

Utility theory is a fascinating subject in its own right, but we do not have time to go into the mathematical details here. Detailed accounts are given by Ferguson (1967) or Berger (1985), for example. Instead, in most of the problems we will be considering, we will use various artificial loss functions. A typical example is use of the loss function

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

the squared error loss function, in a point estimation problem. Then the risk $R(\theta, d)$ of a decision rule is just the mean squared error of $d(X)$ as an estimator of θ, $\mathbb{E}_\theta\{d(X) - \theta\}^2$. In this context, we seek a decision rule d that minimises this mean squared error.

Other commonly used loss functions, in point estimation problems, are

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

the absolute error loss function, and

$$L(\theta, a) = \begin{cases} 0 & \text{if } |\theta - a| \leq \delta, \\ 1 & \text{if } |\theta - a| > \delta, \end{cases}$$

where δ is some prescribed tolerance limit. This latter loss function is useful in a Bayesian formulation of interval estimation, as we shall discuss in Chapter 3.

In hypothesis testing, where we have two hypotheses H_0, H_1, identified with subsets of Θ, and corresponding action space $\mathcal{A} = \{a_0, a_1\}$ in which action a_j corresponds to selecting

the hypothesis H_j, $j = 0, 1$, the most familiar loss function is

$$L(\theta, a) = \begin{cases} 1 & \text{if } \theta \in H_0 \text{ and } a = a_1, \\ 1 & \text{if } \theta \in H_1 \text{ and } a = a_0, \\ 0 & \text{otherwise.} \end{cases}$$

In this case the risk is the probability of making a wrong decision:

$$R(\theta, d) = \begin{cases} \Pr_\theta\{d(X) = a_1\} & \text{if } \theta \in H_0, \\ \Pr_\theta\{d(X) = a_0\} & \text{if } \theta \in H_1. \end{cases}$$

In the classical language of hypothesis testing, these two risks are called, respectively, the type I error and the type II error: see Chapter 4.

2.3 Criteria for a good decision rule

In almost any case of practical interest, there will be no way to find a decision rule $d \in \mathcal{D}$ which makes the risk function $R(\theta, d)$ uniformly smallest for all values of θ. Instead, it is necessary to consider a number of criteria, which help to narrow down the class of decision rules we consider. The notion is to start with a large class of decision rules d, such as the set of *all* functions from $\mathcal{X}$ to $\mathcal{A}$, and then reduce the number of candidate decision rules by application of the various criteria, in the hope of being left with some unique best decision rule for the given inference problem.

2.3.1 Admissibility

Given two decision rules d and d', we say that d *strictly dominates* d' if $R(\theta, d) \leq R(\theta, d')$ for all values of θ, and $R(\theta, d) < R(\theta, d')$ for at least one value θ.

Given a choice between d and d', we would always prefer to use d.

Any decision rule which is strictly dominated by another decision rule (as d' is in the definition) is said to be *inadmissible*. Correspondingly, if a decision rule d is not strictly dominated by any other decision rule, then it is *admissible*.

Admissibility looks like a very weak requirement: it seems obvious that we should always restrict ourselves to admissible decision rules. Admissibility really represents absence of a negative attribute, rather than possession of a positive attribute. In practice, it may not be so easy to decide whether a given decision rule is admissible or not, and there are some surprising examples of natural-looking estimators which are inadmissible. In Chapter 3, we consider an example of an inadmissible estimator, Stein's paradox, which has been described (Efron, 1992) as 'the most striking theorem of post-war mathematical statistics'!

2.3.2 Minimax decision rules

The maximum risk of a decision rule $d \in \mathcal{D}$ is defined by

$$\mathrm{MR}(d) = \sup_{\theta \in \Theta} R(\theta, d).$$

A decision rule d is *minimax* if it minimises the maximum risk:

$$\mathrm{MR}(d) \leq \mathrm{MR}(d') \text{ for all decision rules } d' \in \mathcal{D}.$$

Another way of writing this is to say that d must satisfy

$$\sup_{\theta} R(\theta, d) = \inf_{d' \in \mathcal{D}} \sup_{\theta \in \Theta} R(\theta, d'). \tag{2.1}$$

In most of the problems we will encounter, the supremum and infimum are actually attained, so that we can rewrite (2.1) as

$$\max_{\theta \in \Theta} R(\theta, d) = \min_{d' \in \mathcal{D}} \max_{\theta \in \Theta} R(\theta, d').$$

(Recall that the difference between $\sup_\theta$ and $\max_\theta$ is that the maximum must actually be attained for some $\theta \in \Theta$, whereas a supremum represents a least upper bound that may not actually be attained for any single value of θ. Similarly for infimum and minimum.)

The *minimax principle* says we should use a minimax decision rule.

A few comments about minimaxity are appropriate.

(a) The motivation may be roughly stated as follows: we do not know anything about the true value of θ, therefore we ought to insure ourselves against the worst possible case. There is also an analogy with game theory. In that context, $L(\theta, a)$ represents the penalty suffered by you (as one player in a game) when you choose the action a and your opponent (the other player) chooses θ. If this $L(\theta, a)$ is also the amount gained by your opponent, then this is called a two-person zero-sum game. In game theory, the minimax principle is well established because, in that context, you know that your opponent is trying to choose θ to maximise your loss. See Ferguson (1967) or Berger (1985) for a detailed exposition of the connections between statistical decision theory and game theory.

(b) There are a number of situations in which minimaxity may lead to a counterintuitive result. One situation is when a decision rule d_1 is better than d_2 for all values of θ except in a very small neighbourhood of a particular value, θ_0 say, where d_2 is much better: see Figure 2.1. In this context one might prefer d_1 unless one had particular reason to think that θ_0, or something near it, was the true parameter value. From a slightly broader perspective, it might seem illogical that the minimax criterion's preference for d_2 is based entirely in its behaviour in a small region of Θ, while the rest of the parameter space is ignored.

(c) The minimax procedure may be likened to an arms race in which both sides spend the maximum sum available on military fortification in order to protect themselves against the worst possible outcome, of being defeated in a war, an instance of a non-zero-sum game!

(d) Minimax rules may not be unique, and may not be admissible. Figure 2.2 is intended to illustrate a situation in which d_1 and d_2 achieve the same minimax risk, but one would obviously prefer d_1 in practice.

2.3.3 Unbiasedness

A decision rule d is said to be *unbiased* if

$$\mathbb{E}_\theta\{L(\theta', d(X))\} \geq \mathbb{E}_\theta\{L(\theta, d(X))\} \text{ for all } \theta, \theta' \in \Theta.$$

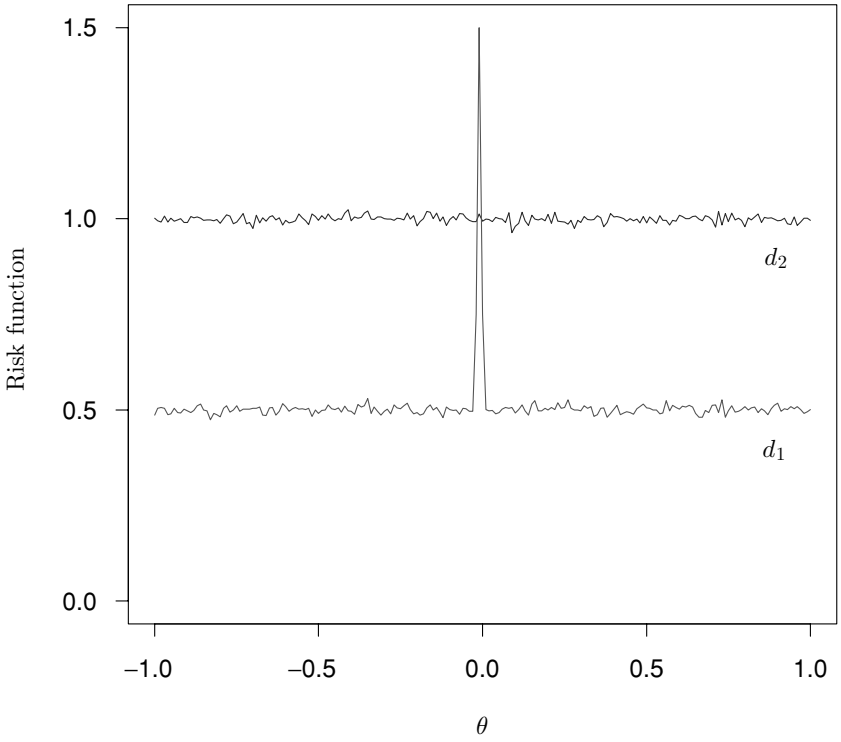

Figure 2.1 Risk functions for two decision rules

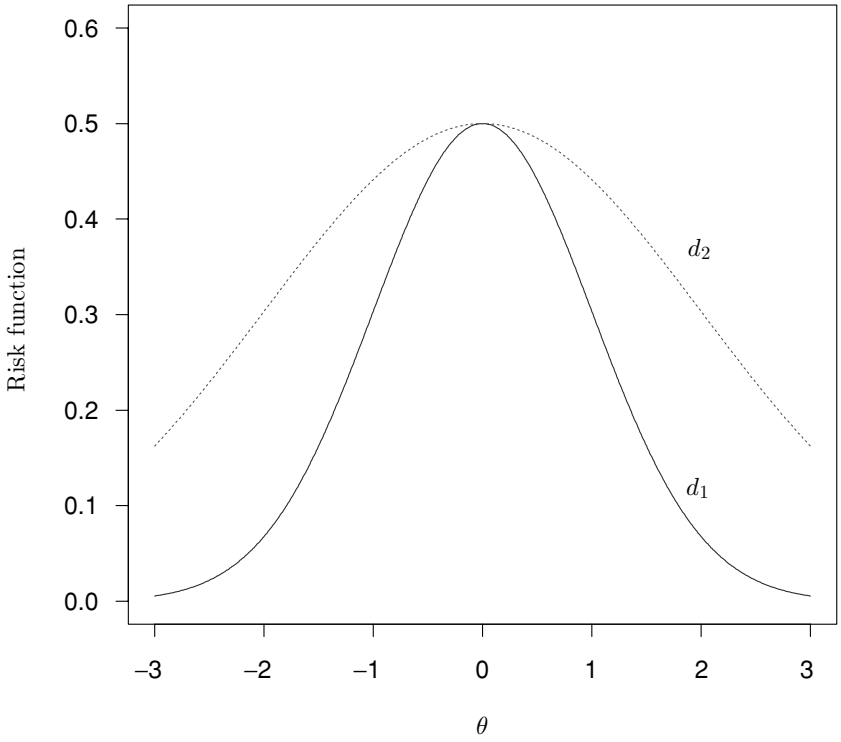

Figure 2.2 Minimax rules may not be admissible

Recall that in elementary statistical theory, if $d(X)$ is an estimator for a parameter θ, then $d(X)$ is said to be unbiased if $\mathbb{E}_\theta d(X) = \theta$ for all θ. The connection between the two notions of unbiasedness is as follows. Suppose the loss function is the squared error loss, $L(\theta, d) = (\theta - d)^2$. Fix θ and let $\mathbb{E}_\theta d(X) = \phi$. Then, for d to be an unbiased decision rule, we require that, for all θ',

$$\begin{aligned} 0 \leq \mathbb{E}_\theta\{L(\theta', d(X))\} - \mathbb{E}_\theta\{L(\theta, d(X))\} &= \mathbb{E}_\theta\{(\theta' - d(X))^2\} - \mathbb{E}_\theta\{(\theta - d(X))^2\} \\ &= (\theta')^2 - 2\theta'\phi + \mathbb{E}_\theta d^2(X) - \theta^2 \\ &\quad + 2\theta\phi - \mathbb{E}_\theta d^2(X) \\ &= (\theta' - \phi)^2 - (\theta - \phi)^2. \end{aligned}$$

If $\phi = \theta$, then this statement is obviously true. If $\phi \neq \theta$, then set $\theta' = \phi$ to obtain a contradiction.

Thus we see that, if $d(X)$ is an unbiased estimator in the classical sense, then it is also an unbiased decision rule, provided the loss is a squared error. However the above argument also shows that the notion of an unbiased decision rule is much broader: we could have whole families of unbiased decision rules corresponding to different loss functions.

Nevertheless, the role of unbiasedness in statistical decision theory is ambiguous. Of the various criteria being considered here, it is the only one that does not depend solely on the risk function. Often we find that biased estimators perform better than unbiased ones from the point of view of, say, minimising mean squared error. For this reason, many modern statisticians consider the whole concept of unbiasedness to be somewhere between a distraction and a total irrelevance.

2.3.4 Bayes decision rules

Bayes decision rules are based on different assumptions from the other criteria we have considered, because, in addition to the loss function and the class $\mathcal{D}$ of decision rules, we must specify a *prior distribution*, which represents our prior knowledge on the value of the parameter θ, and is represented by a function $\pi(\theta),\ \theta \in \Theta$. In cases where Θ contains an open rectangle in $\mathbb{R}^d$, we would take our prior distribution to be absolutely continuous, meaning that $\pi(\theta)$ is taken to be some probability density on Θ. In the case of a discrete parameter space, $\pi(\theta)$ is a probability mass function.

In the continuous case, the Bayes risk of a decision rule d is defined to be

$$r(\pi, d) = \int_{\theta \in \Theta} R(\theta, d)\pi(\theta)d\theta.$$

In the discrete case, the integral in this expression is replaced by a summation over the possible values of θ. So, the Bayes risk is just average risk, the averaging being with respect to the weight function $\pi(\theta)$ implied by our prior distribution.

A decision rule d is said to be a *Bayes rule*, with respect to a given prior $\pi(\cdot)$, if it minimises the Bayes risk, so that

$$r(\pi, d) = \inf_{d' \in \mathcal{D}} r(\pi, d') = m_\pi, \text{ say.} \tag{2.2}$$

The *Bayes principle* says we should use a Bayes decision rule.

2.3.5 Some other definitions

Sometimes the Bayes rule is not defined because the infimum in (2.2) is not attained for any decision rule d. However, in such cases, for any $\epsilon > 0$ we can find a decision rule d_ϵ for which

$$r(\pi, d_\epsilon) < m_\pi + \epsilon$$

and in this case d_ϵ is said to be *ϵ-Bayes* with respect to the prior distribution $\pi(\cdot)$.

Finally, a decision rule d is said to be *extended Bayes* if, for every $\epsilon > 0$, we have that d is ϵ-Bayes with respect to *some* prior, which need not be the same for different ϵ. As we shall see in Theorem 2.2, it is often possible to derive a minimax rule through the property of being extended Bayes. A particular example of an extended Bayes rule is discussed in Problem 3.11.

2.4 Randomised decision rules

Suppose we have a collection of I decision rules $d_1, \ldots, d_I$ and an associated set of probability weights $p_1, \ldots, p_I$, so that $p_i \geq 0$ for $1 \leq i \leq I$, and $\sum_i p_i = 1$.

Define the decision rule $d^* = \sum_i p_i d_i$ to be the rule 'select d_i with probability p_i'. Then d^* is a *randomised decision rule*. We can imagine that we first use some randomisation mechanism, such as tossing coins or using a computer random number generator, to select, independently of the observed data x, one of the decision rules $d_1, \ldots, d_I$, with respective probabilities $p_1, \ldots, p_I$. Then, having decided in favour of use of the particular rule d_i, under d^* we carry out the action $d_i(x)$.

For a randomised decision rule d^*, the risk function is defined by averaging across possible risks associated with the component decision rules:

$$R(\theta, d^*) = \sum_{i=1}^{I} p_i R(\theta, d_i).$$

Randomised decision rules may appear to be artificial, but minimax solutions may well be of this form. It is easy to contruct examples in which d^* is formed by randomising the rules $d_1, \ldots, d_I$ but

$$\sup_\theta R(\theta, d^*) < \sup_\theta R(\theta, d_i) \text{ for each } i,$$

so that d^* may be a candidate for the minimax procedure, but none of $d_1, \ldots, d_I$. An example of a decision problem, where the minimax rule indeed turns out to be a randomised rule, is presented in Section 2.5.1, and illustrated in Figure 2.9.

2.5 Finite decision problems

A finite decision problem is one in which the *parameter space* is a finite set: $\Theta = \{\theta_1, \ldots, \theta_t\}$ for some finite t, with $\theta_1, \ldots, \theta_t$ specified values. In such cases the notions of admissible, minimax and Bayes decision rules can be given a geometric interpretation, which leads to some interesting problems in their own right, and which also serves to motivate some properties of decision rules in more general problems.

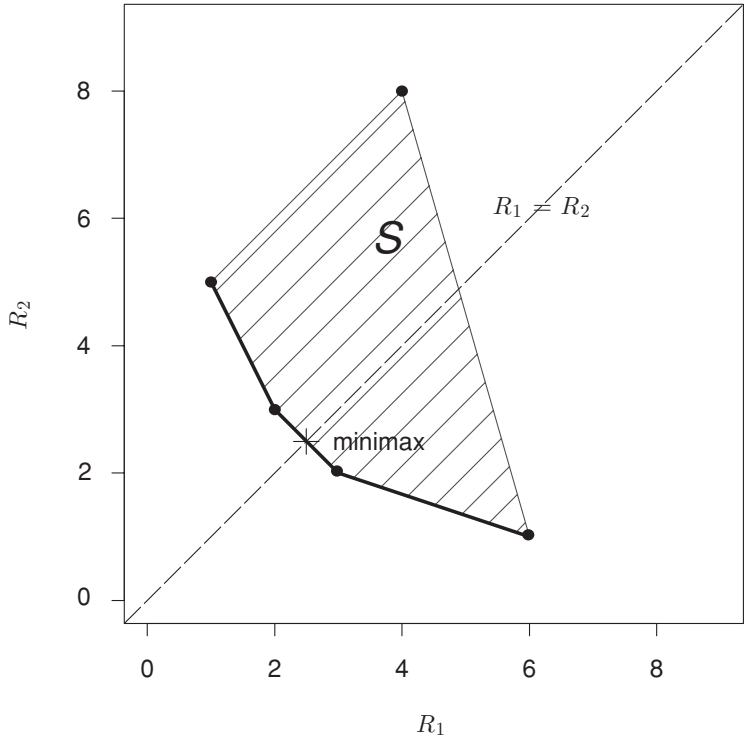

Figure 2.3 An example of a risk set

For a finite decision problem, define the *risk set* to be a subset S of $\mathbb{R}^t$, in which a generic point consists of the t-vector $(R(\theta_1, d), \ldots, R(\theta_t, d))$ associated with a decision rule d. An important point to note is that we assume in our subsequent discussion that the space of decision rules $\mathcal{D}$ includes all randomised rules.

A set A is said to be *convex* if whenever $x_1 \in A$ and $x_2 \in A$ then $\lambda x_1 + (1 - \lambda)x_2 \in A$ for any $\lambda \in (0, 1)$.

Lemma 2.1 *The risk set S is a convex set.*

Proof Suppose $x_1 = (R(\theta_1, d_1), \ldots, R(\theta_t, d_1))$ and $x_2 = (R(\theta_1, d_2), \ldots, R(\theta_t, d_2))$ are two elements of S, and suppose $\lambda \in (0, 1)$. Form a new randomised decision rule $d = \lambda d_1 + (1 - \lambda)d_2$. Then for every θ, by definition of the risk of a randomised rule,

$$R(\theta, d) = \lambda R(\theta, d_1) + (1 - \lambda)R(\theta, d_2).$$

Then we see that $\lambda x_1 + (1 - \lambda)x_2$ is associated with the decision rule d, and hence is itself a member of S. This proves the result. □

In the case $t = 2$, it is particularly easy to see what is going on, because we can draw the risk set as a subset of $\mathbb{R}^2$, with coordinate axes $R_1 = R(\theta_1, d)$, $R_2 = R(\theta_2, d)$. An example is shown in Figure 2.3.

The extreme points of S (shown by the dots) correspond to non-randomised decision rules, and points on the lower left-hand boundary (represented by thicker lines) correspond to the admissible decision rules.

For the example shown in Figure 2.3, the minimax decision rule corresponds to the point at the lower intersection of S with the line $R_1 = R_2$ (the point shown by the cross). Note

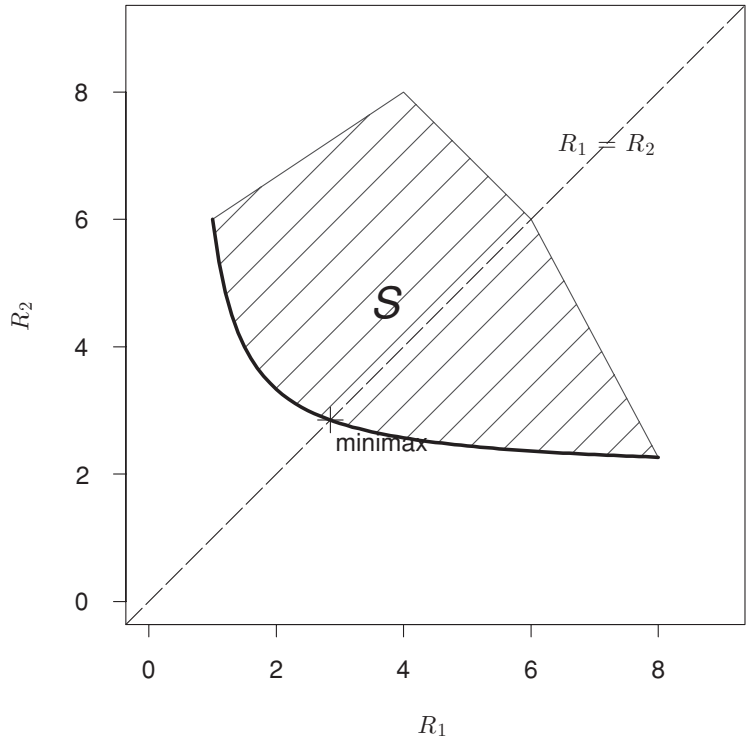

Figure 2.4 Another risk set

that this is a randomised decision rule. However, we shall see below that not all minimax rules are formed in this way.

In cases where the boundary of the risk set is a smooth curve, as in Figure 2.4, the properties are similar, except that we now have an infinite number of non-randomised admissible decision rules and the minimax point is non-randomised.

To find the Bayes rules, suppose we have prior probabilities (π_1, π_2), where $\pi_1 \geq 0, \pi_2 \geq 0, \pi_1 + \pi_2 = 1$, so that π_j, for $j = 1$ or 2, represents the prior probability that θ_j is the true parameter value. For any c, the straight line $\pi_1 R_1 + \pi_2 R_2 = c$ represents a class of decision rules with the same Bayes risk. By varying c, we get a family of parallel straight lines. Of course, if the line $\pi_1 R_1 + \pi_2 R_2 = c$ does not intersect S, then the Bayes risk c is unattainable. Then the Bayes risk for the decision problem is c' if the line $\pi_1 R_1 + \pi_2 R_2 = c'$ just hits the set S on its lower left-hand boundary: see Figure 2.5. Provided S is a closed set, which we shall assume, the Bayes decision rule corresponds to the point at which this line intersects S.

In many cases of interest, the Bayes rule is unique, and is then automatically both admissible and non-randomised. However, it is possible that the line $\pi_1 R_1 + \pi_2 R_2 = c'$ hits S along a line segment rather than at a single point, as in Figure 2.6. In that case, any point along the segment identifies a Bayes decision rule with respect to this prior. Also, in this case the interior points of the segment will identify randomised decision rules, but the endpoints of the segment also yield Bayes rules, which are non-randomised. Thus we can always find a Bayes rule which is non-randomised. Also, it is an easy exercise to see that, provided $\pi_1 > 0, \pi_2 > 0$, the Bayes rule is admissible.

It can easily happen (see Figure 2.7) that the same decision rule is Bayes with respect to a whole family of different prior distributions.

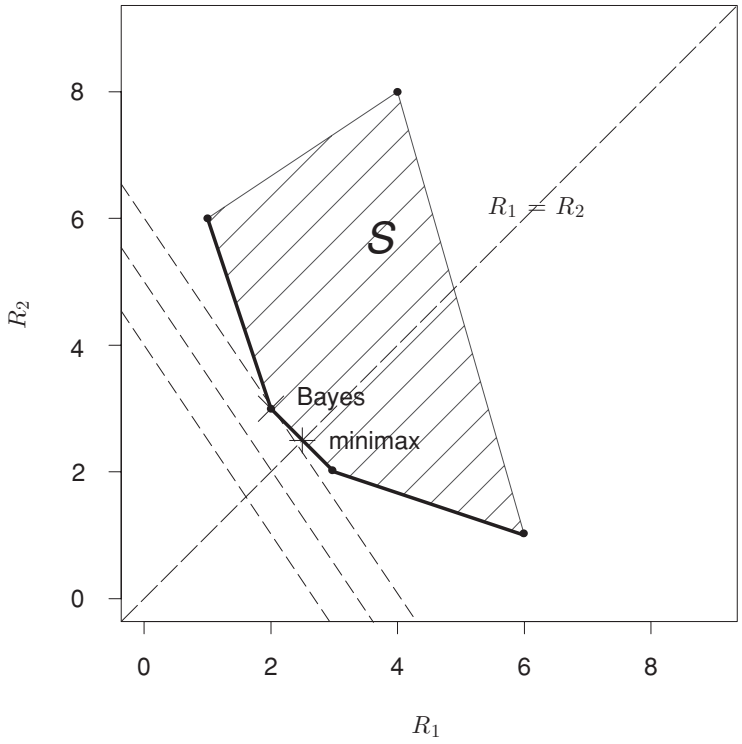

Figure 2.5 Finding the Bayes rule

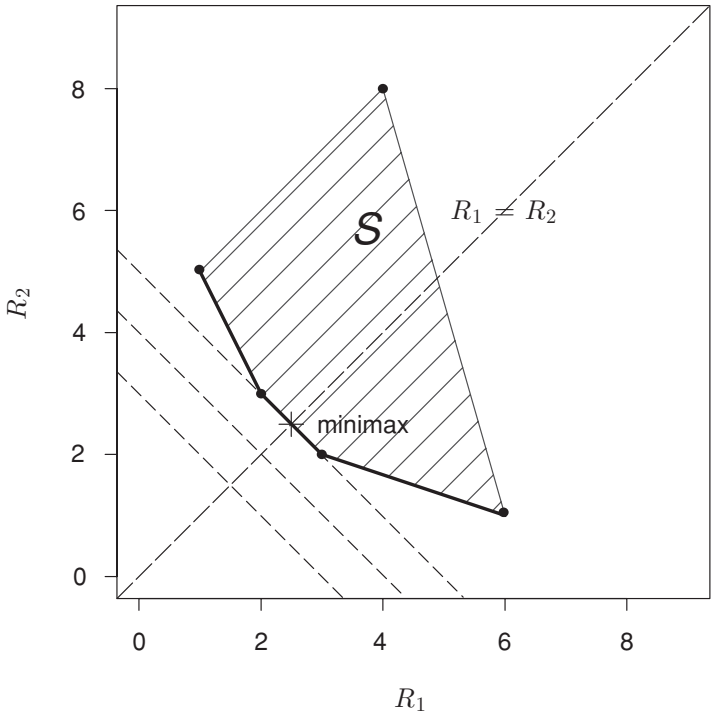

Figure 2.6 Non-uniqueness of Bayes rule

Not all minimax rules satisfy $R_1 = R_2$. See Figures 2.8 for several examples. By contrast, Figure 2.3 is an example where the minimax rule does satisfy $R_1 = R_2$. In Figures 2.8(a) and (b), S lies entirely to the left of the line $R_1 = R_2$, so that $R_1 < R_2$ for every point in S, and therefore the minimax rule is simply that which minimises R_2. This is the

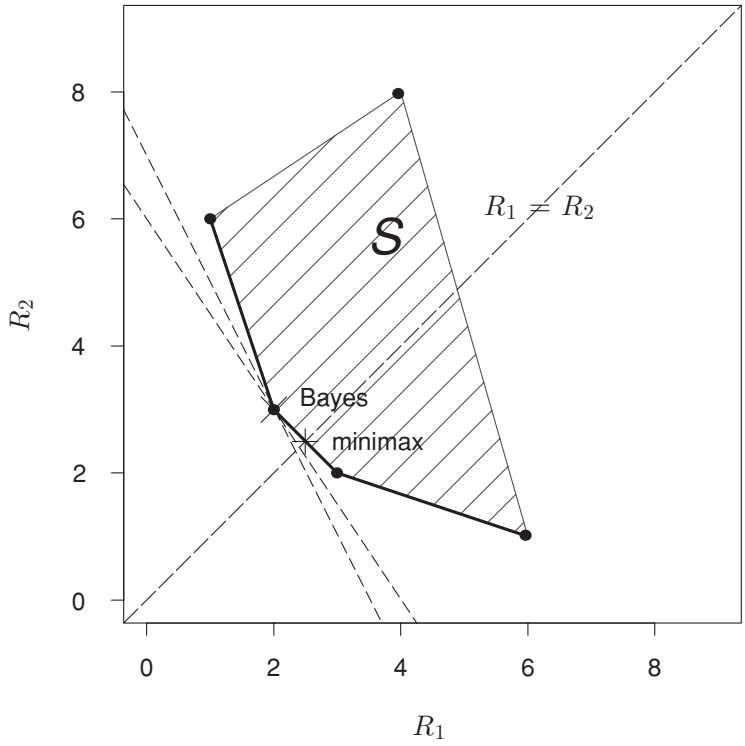

Figure 2.7 Bayes with respect to family of priors

Bayes rule for the prior $\pi_1 = 0, \pi_2 = 1$, and may be attained either at a single point (Figure 2.8(a)) or along a segment (Figure 2.8(b)). In the latter case, only the left-hand element of the segment (shown by the cross) corresponds to an admissible rule. Figures 2.8(c) and (d) are the mirror images of Figures 2.8(a) and (b) in which every point of the risk set satisfies $R_1 > R_2$ so that the minimax rule is Bayes for the prior $\pi_1 = 1$, $\pi_2 = 0$.

2.5.1 A story

The Palliser emerald necklace has returned from the cleaners, together with a valueless imitation which you, as Duchess of Omnium, wear on the less important State occasions. The tag identifying the imitation has fallen off, and so you have two apparently identical necklaces in the left- and right-hand drawers of the jewelcase. You consult your Great Aunt, who inspects them both (left-hand necklace first, and then right-hand necklace), and then from her long experience pronounces one of them to be the true necklace. *But is she right?* You know that her judgement will be infallible if she happens to inspect the true necklace first and the imitation afterwards, but that if she inspects them in the other order she will in effect select one of them at random, with equal probabilities $\frac{1}{2}$ on the two possibilities. With a loss of £0 being attributed to a correct decision (choosing the real necklace), you know that a mistaken one (choosing the imitation) will imply a loss of £1 million. You will wear the necklace tonight at an important banquet, where the guest of honour is not only the Head of State of a country with important business contracts with Omnium, but also an expert on emerald jewellery, certain to be able to spot an imitation.

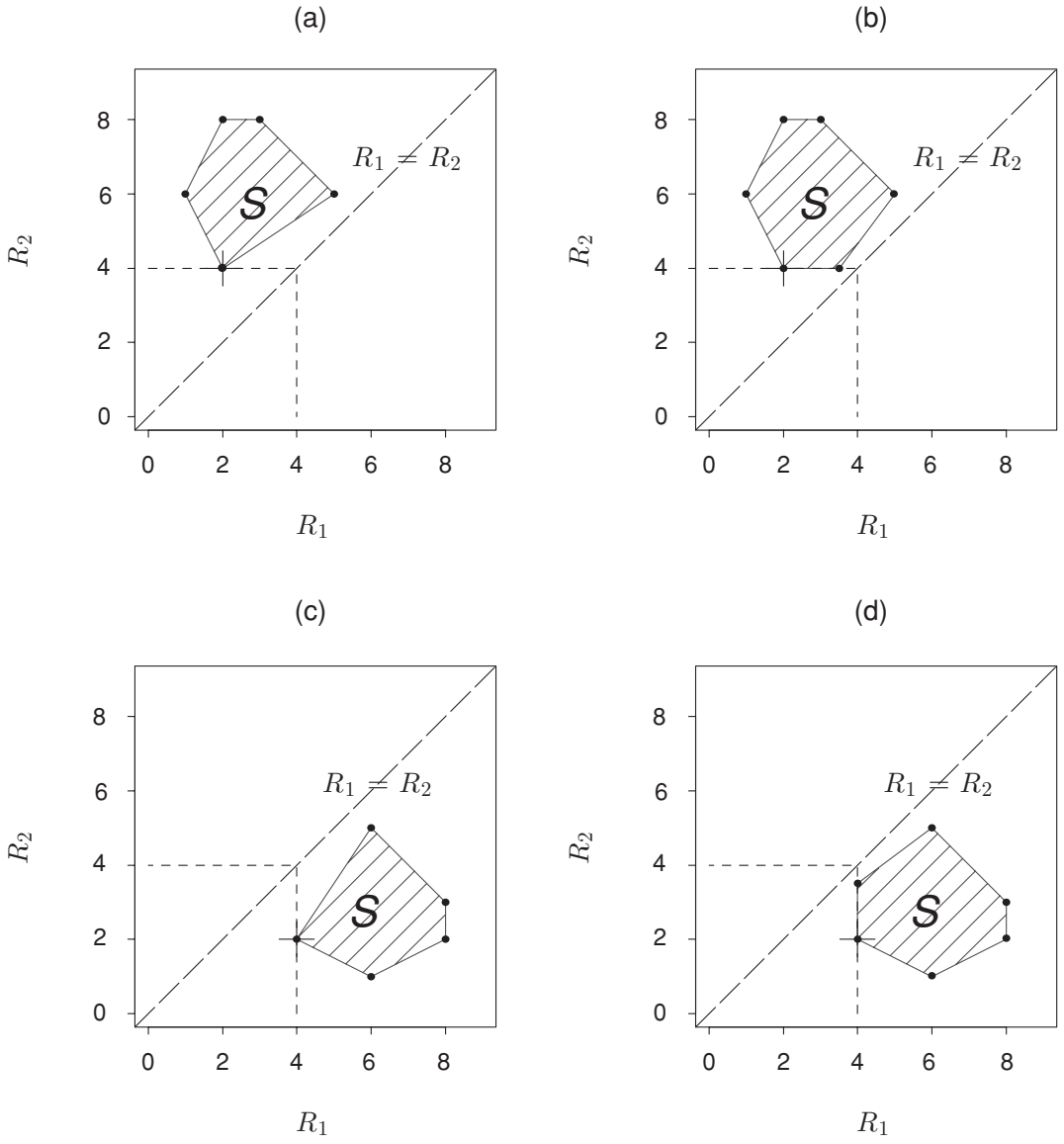

Figure 2.8 Possible forms of risk set

Our data x are your Great Aunt's judgement. Consider the following exhaustive set of four non-randomised decision rules based on x:

(d_1): accept the left-hand necklace, irrespective of Great Aunt's judgement;
(d_2): accept the right-hand necklace, irrespective of Great Aunt's judgement;
(d_3): accept your Great Aunt's judgement;
(d_4): accept the reverse of your Great Aunt's judgement.

Code the states of nature 'left-hand necklace is the true one' as $\theta = 1$, and 'right-hand necklace is the true one' as $\theta = 2$. We compute the risk functions of the decision rules as follows:

$R_1 = R(\theta = 1, d_1) = 0$, $R_2 = R(\theta = 2, d_1) = 1$;
$R_1 = R(\theta = 1, d_2) = 1$, $R_2 = R(\theta = 2, d_2) = 0$;
$R_1 = R(\theta = 1, d_3) = 0$, $R_2 = R(\theta = 2, d_3) = \frac{1}{2}$;
$R_1 = R(\theta = 1, d_4) = 1$, $R_2 = R(\theta = 2, d_4) = \frac{1}{2}$.

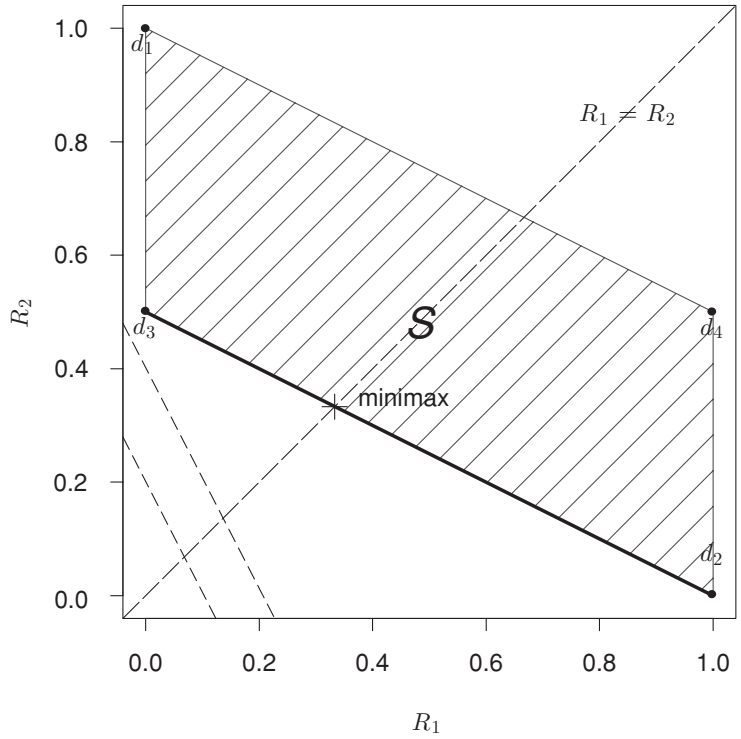

Figure 2.9 Risk set, Palliser necklace

To understand these risks, note that when $\theta = 1$ your Great Aunt chooses correctly, so that d_3 is certain to make the correct choice, while d_4 is certain to make the wrong choice. When $\theta = 2$, Great Aunt chooses incorrectly with probability $\frac{1}{2}$, so our expected loss is $\frac{1}{2}$ with both rules d_3 and d_4.

The risks associated with these decision rules and the associated risk set are shown in Figure 2.9. The only admissible rules are d_2 and d_3, together with randomised rules formed as convex combinations of d_2 and d_3.

The minimax rule d^* is such a randomised rule, of the form $d^* = \lambda d_3 + (1 - \lambda)d_2$. Then setting $R(\theta = 1, d^*) = R(\theta = 2, d^*)$ gives $\lambda = \frac{2}{3}$.

Suppose that the Duke now returns from hunting and points out that the jewel cleaner will have placed the true necklace in the left-hand drawer with some probability ψ, which we know from knowledge of the way the jewelcase was arranged on its return from previous trips to the cleaners.

The Bayes risk of a rule d is then $\psi R(\theta = 1, d) + (1 - \psi)R(\theta = 2, d)$. There are three groups of Bayes rules according to the value of ψ:

(i) If $\psi = \frac{1}{3}$, d_2 and d_3, together with all convex combinations of the two, give the same Bayes risk ($= \frac{1}{3}$) and are Bayes rules. In this case the minimax rule is also a Bayes rule.

(ii) If $\psi > \frac{1}{3}$ the unique Bayes rule is d_3, with Bayes risk $(1 - \psi)/2$. This is the situation with the lines of constant Bayes risk illustrated in Figure 2.9, and makes intuitive sense. If our prior belief that the true necklace is in the left-hand drawer is strong, then we attach a high probability to Great Aunt inspecting the necklaces in the order true necklace first, then imitation, in which circumstances she is certain to identify them correctly. Then following her judgement is sensible.

(iii) If $\psi < \frac{1}{3}$ the unique Bayes rule is d_2, with Bayes risk ψ. Now the prior belief is that the true necklace is unlikely to be in the left-hand drawer, so we are most likely to be

in the situation where Great Aunt basically guesses which is the true necklace, and it is better to go with our prior hunch of the right-hand drawer.

2.6 Finding minimax rules in general

Although the formula $R_1 = R_2$ to define the minimax rule is satisfied in situations such as that shown in Figure 2.3, all four situations illustrated in Figure 2.8. satisfy a more general formula:

$$\max_{\theta} R(\theta, d) \leq r(\pi, d), \tag{2.3}$$

where π is *a* prior distribution with respect to which the minimax rule d is Bayes. In the cases of Figure 2.8(a) and Figure 2.8(b), for example, this is true for the minimax rules, for the prior $\pi_1 = 0, \pi_2 = 1$, though we note that in Figure 2.8(a) the unique minimax rule is Bayes for other priors as well.

These geometric arguments suggest that, in general, a minimax rule is one which satisfies:

(a) it is Bayes with respect to some prior $\pi(\cdot)$,
(b) it satisfies (2.3).

A complete classification of minimax decision rules in general problems lies outside the scope of this text, but the following two theorems give simple sufficient conditions for a decision rule to be minimax. One generalisation that is needed in passing from the finite to the infinite case is that the class of Bayes rules must be extended to include sequences of either Bayes rules, or extended Bayes rules.

Theorem 2.1 *If δ_n is Bayes with respect to prior $\pi_n(\cdot)$, and $r(\pi_n, \delta_n) \to C$ as $n \to \infty$, and if $R(\theta, \delta_0) \leq C$ for all $\theta \in \Theta$, then δ_0 is minimax.*

Of course this includes the case where $\delta_n = \delta_0$ for all n and the Bayes risk of δ_0 is exactly C.

To see the infinite-dimensional generalisation of the condition $R_1 = R_2$, we make the following definition.

Definition *A decision rule d is an* equaliser decision rule *if $R(\theta, d)$ is the same for every value of θ.*

Theorem 2.2 *An equaliser decision rule δ_0 which is extended Bayes must be minimax.*

Proof of Theorem 2.1 Suppose δ_0 satisfies the conditions of the theorem but is not minimax. Then there must exist some decision rule δ' for which $\sup_\theta R(\theta, \delta') < C$: the inequality must be strict, because, if the maximum risk of δ' was the same as that of δ_0, that would not contradict minimaxity of δ_0. So there is an $\epsilon > 0$ for which $R(\theta, \delta') < C - \epsilon$ for every θ. Now, since $r(\pi_n, \delta_n) \to C$, we can find an n for which $r(\pi_n, \delta_n) > C - \epsilon/2$. But $r(\pi_n, \delta') \leq C - \epsilon$. Therefore, δ_n cannot be the Bayes rule with respect to π_n. This creates a contradiction, and hence proves the theorem. □

Proof of Theorem 2.2. The proof here is almost the same. If we suppose δ_0 is not minimax, then there exists a δ' for which $\sup_\theta R(\theta, \delta') < C$, where C is the common value of $R(\theta, \delta_0)$.

So let $\sup_\theta R(\theta, \delta') = C - \epsilon$, for some $\epsilon > 0$. By the extended Bayes property of δ_0, we can find a prior π for which

$$r(\pi, \delta_0) = C < \inf_\delta r(\pi, \delta) + \frac{\epsilon}{2}.$$

But $r(\pi, \delta') \leq C - \epsilon$, so this gives another contradiction, and hence proves the theorem. □

2.7 Admissibility of Bayes rules

In Chapter 3 we will present a general result that allows us to characterise the Bayes decision rule for any given inference problem. An immediate question that then arises concerns admissibility. In that regard, the rule of thumb is that Bayes rules are nearly always admissible. We complete this chapter with some specific theorems on this point. Proofs are left as exercises.

Theorem 2.3 *Assume that $\Theta = \{\theta_1, \ldots, \theta_t\}$ is finite, and that the prior $\pi(\cdot)$ gives positive probability to each θ_i. Then a Bayes rule with respect to π is admissible.*

Theorem 2.4 *If a Bayes rule is unique, it is admissible.*

Theorem 2.5 *Let Θ be a subset of the real line. Assume that the risk functions $R(\theta, d)$ are continuous in θ for all decision rules d. Suppose that for any $\epsilon > 0$ and any θ the interval $(\theta - \epsilon, \theta + \epsilon)$ has positive probability under the prior $\pi(\cdot)$. Then a Bayes rule with respect to π is admissible.*