1

Introduction

What is statistical inference?

In statistical inference experimental or observational data are modelled as the observed values of random variables, to provide a framework from which inductive conclusions may be drawn about the mechanism giving rise to the data.

We wish to analyse observations $x = (x_1, \ldots, x_n)$ by:

1 Regarding x as the observed value of a random variable $X = (X_1, \ldots, X_n)$ having an (unknown) probability distribution, conveniently specified by a probability density, or probability mass function, $f(x)$.
2 Restricting the unknown density to a suitable family or set $\mathcal{F}$. In *parametric statistical inference*, $f(x)$ is of known analytic form, but involves a finite number of real unknown parameters $\theta = (\theta_1, \ldots, \theta_d)$. We specify the region $\Theta \subseteq \mathbb{R}^d$ of possible values of θ, the *parameter space*. To denote the dependency of $f(x)$ on θ, we write $f(x; \theta)$ and refer to this as the *model function*. Alternatively, the data could be modelled non-parametrically, a non-parametric model simply being one which does not admit a parametric representation. We will be concerned almost entirely in this book with parametric statistical inference.

The objective that we then assume is that of assessing, on the basis of the observed data x, some aspect of θ, which for the purpose of the discussion in this paragraph we take to be the value of a particular component, θ_i say. In that regard, we identify three main types of inference: *point estimation*, *confidence set estimation* and *hypothesis testing*. In point estimation, a single value is computed from the data x and used as an estimate of θ_i. In confidence set estimation we provide a set of values, which, it is hoped, has a predetermined high probability of including the true, but unknown, value of θ_i. Hypothesis testing sets up specific hypotheses regarding θ_i and assesses the plausibility of any such hypothesis by assessing whether or not the data x support that hypothesis.

Of course, other objectives might be considered, such as: (a) prediction of the value of some as yet unobserved random variable whose distribution depends on θ, or (b) examination of the adequacy of the model specified by $\mathcal{F}$ and Θ. These are important problems, but are not the main focus of the present book, though we will say a little on predictive inference in Chapter 10.

How do we approach statistical inference?

Following Efron (1998), we identify three main paradigms of statistical inference: the *Bayesian, Fisherian* and *frequentist*. A key objective of this book is to develop in detail the essential features of all three schools of thought and to highlight, we hope in an interesting way, the potential conflicts between them. The basic differences that emerge relate to interpretation of probability and to the objectives of statistical inference. To set the scene, it is of some value to sketch straight away the main characteristics of the three paradigms. To do so, it is instructive to look a little at the historical development of the subject.

The Bayesian paradigm goes back to Bayes and Laplace, in the late eighteenth century. The fundamental idea behind this approach is that the unknown parameter, θ, should itself be treated as a random variable. Key to the Bayesian viewpoint, therefore, is the specification of a *prior probability distribution* on θ, before the data analysis. We will describe in some detail in Chapter 3 the main approaches to specification of prior distributions, but this can basically be done either in some objective way, or in a subjective way, which reflects the statistician's own prior state of belief. To the Bayesian, inference is the formalisation of how the prior distribution changes, to the *posterior distribution*, in the light of the evidence presented by the available data x, through Bayes' formula. Central to the Bayesian perspective, therefore, is a use of probability distributions as expressing opinion.

In the early 1920s, R.A. Fisher put forward an opposing viewpoint, that statistical inference must be based entirely on probabilities with direct experimental interpretation. As Efron (1998) notes, Fisher's primary concern was the development of a logic of inductive inference, which would release the statistician from the a priori assumptions of the Bayesian school. Central to the Fisherian viewpoint is the *repeated sampling principle*. This dictates that the inference we draw from x should be founded on an analysis of how the conclusions change with variations in the data samples, which would be obtained through hypothetical repetitions, under exactly the same conditions, of the experiment which generated the data x in the first place. In a Fisherian approach to inference, a central role is played by the concept of *likelihood*, and the associated principle of *maximum likelihood*. In essence, the likelihood measures the probability that different values of the parameter θ assign, under a hypothetical repetition of the experiment, to re-observation of the actual data x. More formally, the ratio of the likelihood at two different values of θ compares the relative plausibilities of observing the data x under the models defined by the two θ values. A further fundamental element of Fisher's viewpoint is that inference, in order to be as relevant as possible to the data x, must be carried out *conditional* on everything that is known and uninformative about θ.

Fisher's greatest contribution was to provide for the first time an optimality yardstick for statistical estimation, a description of the optimum that it is possible to do in a given estimation problem, and the technique of maximum likelihood, which produces estimators of θ that are close to ideal in terms of that yardstick. As described by Pace and Salvan (1997), spurred on by Fisher's introduction of optimality ideas in the 1930s and 1940s, Neyman, E.S. Pearson and, later, Wald and Lehmann offered the third of the three paradigms, the frequentist approach. The origins of this approach lay in a detailed mathematical analysis of some of the fundamental concepts developed by Fisher, in particular likelihood and *sufficiency*. With this focus, emphasis shifted from inference as a summary of data, as

favoured by Fisher, to inferential procedures viewed as decision problems. Key elements of the frequentist approach are the need for clarity in mathematical formulation, and that optimum inference procedures should be identified *before* the observations x are available, optimality being defined explicitly in terms of the repeated sampling principle.

The plan of the book is as follows. In Chapter 2, we describe the main elements of the *decision theory* approach to frequentist inference, where a strict mathematical statement of the inference problem is made, followed by formal identification of the optimal solution. Chapter 3 develops the key ideas of Bayesian inference, before we consider, in Chapter 4, central optimality results for hypothesis testing from a frequentist perspective. There a comparison is made between frequentist and Bayesian approaches to hypothesis testing. Chapter 5 introduces two special classes of model function of particular importance to later chapters, exponential families and transformation models. Chapter 6 is concerned primarily with point estimation of a parameter θ and provides a formal introduction to a number of key concepts in statistical inference, in particular the notion of sufficiency. In Chapter 7, we revisit the topic of hypothesis testing, to extend some of the ideas of Chapter 4 to more complicated settings. In the former chapter we consider also the conditionality ideas that are central to the Fisherian perspective, and highlight conflicts with the frequentist approach to inference. There we describe also key frequentist optimality ideas in confidence set estimation. The subject of Chapter 8 is maximum likelihood and associated inference procedures. The remaining chapters contain more advanced material. In Chapter 9 we present a description of some recent innovations in statistical inference, concentrating on ideas which draw their inspiration primarily from the Fisherian viewpoint. Chapter 10 provides a discussion of various approaches to *predictive inference*. Chapter 11, reflecting the personal interests of one of us (GAY), provides a description of the *bootstrap* approach to inference. This approach, made possible by the recent availability of cheap computing power, offers the prospect of techniques of statistical inference which avoid the need for awkward mathematical analysis, but retain the key operational properties of methods of inference studied elsewhere in the book, in particular in relation to the repeated sampling principle.

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

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

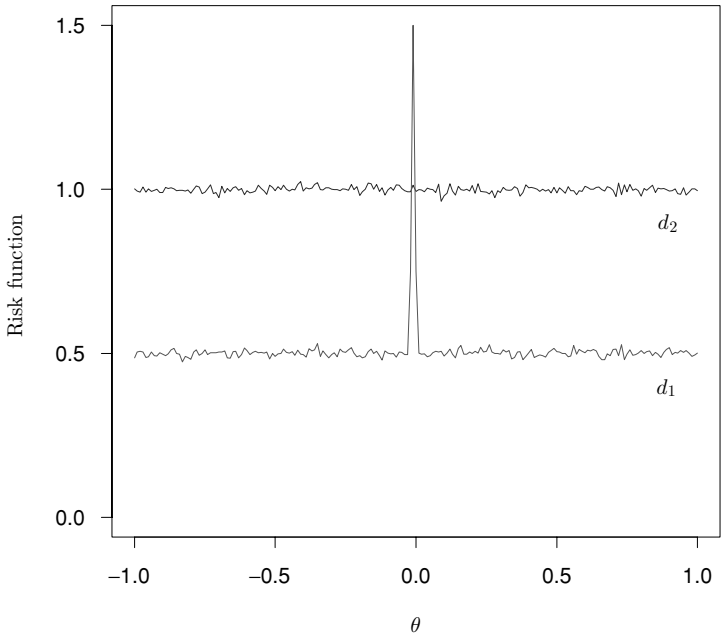

Figure 2.1 Risk functions for two decision rules

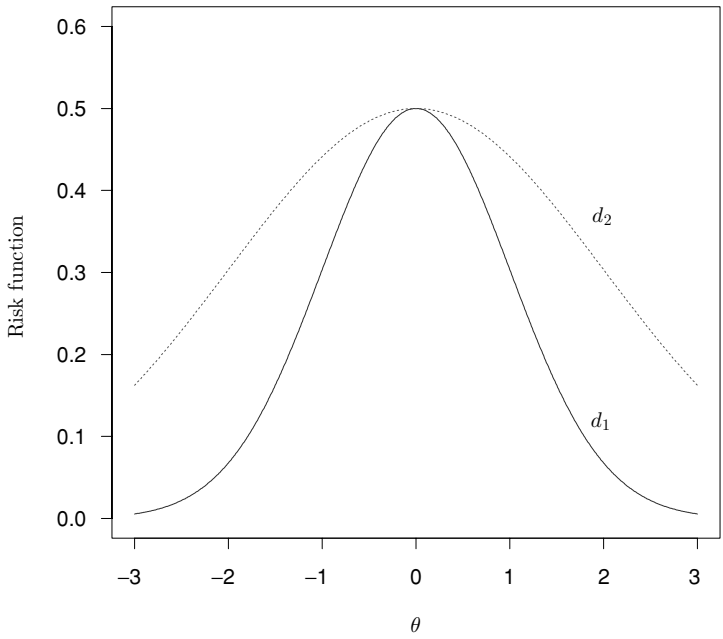

Figure 2.2 Minimax rules may not be admissible

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',

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

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.