CHAPTER 2

Decision-Theoretic Foundations

> Today would run out according to the Pattern. But over and over he mulled over the decisions he had made since he first entered the Waste. Could he have done something different, something that would have avoided this day, this place? Next time, perhaps.
>
> **Robert Jordan**, *The Fires of Heaven, Book V of the Wheel of Time.*

2.1 Evaluating estimators

Considering that the overall purpose of most inferential studies is to provide the statistician (or a client) with a *decision*, it seems reasonable to ask for an *evaluation* criterion of decision procedures that assesses the consequences of each decision and depends on the parameters of the model, i.e., the true state of the world (or of Nature). These *decisions* can be of various kinds, ranging from buying equities depending on their future returns θ, to stopping an agricultural experiment on the productivity θ of a new crop species, to estimating the underground economy contribution θ to the U.S. GNP, to deciding whether the number θ of homeless people has increased since the last census. They also include assessing whether a new scientific theory is compatible with the experimental evidence at hand. If no evaluation criterion is available, it is impossible to compare different decision procedures and absurd solutions, such as proposing $\hat{\theta} = 3$ for any real estimation problem or even more dramatically the answer one wants to impose, can only be eliminated by ad-hoc reasoning. To avoid such reasoning implies a reinforced axiomatization of the statistical inferential framework, called *Decision Theory*. This augmented theoretical structure is necessary for Statistics to reach a coherence otherwise unattainable[1].

Although almost everybody agrees on the need for such an evaluation criterion, there is an important controversy running about the choice of this

[1] The Bayesian approach is, from our point of view, the ultimate step in this quest for coherence.

evaluation criterion, since the consequences on the decision are not innocuous. This difficulty even led some statisticians to totally reject Decision Theory, on the basis that a practical determination of the decision-maker evaluation criterion is utterly impossible in most cases.

This criterion is usually called *loss* and is defined as follows, where $\mathcal{D}$ denotes the set of possible decisions. $\mathcal{D}$ is called the *decision* space and most theoretical examples focus on the case $\mathcal{D} = \Theta$, which represents the standard estimation setting.

Definition 2.1.1 *A loss function is any function* L *from* $\Theta \times \mathcal{D}$ *in* $[0, +\infty)$.

This loss function is supposed to evaluate the penalty (or error) $\mathrm{L}(\theta, d)$ associated with the decision d when the parameter takes the value θ. In a traditional setting of parameter estimation, when $\mathcal{D}$ is Θ or $h(\Theta)$, the loss function $\mathrm{L}(\theta, \delta)$ measures the error made in evaluating $h(\theta)$ by δ. Section 2.2 introduces a set of so-called rationality axioms that ensures the existence of such a function in a decision setting.

The actual determination of the loss function is often awkward in practice, in particular because the determination of the consequences of each action for each value of θ is usually impossible when $\mathcal{D}$ or Θ are large sets, for instance when they have an infinite number of elements. Moreover, in qualitative models, it may be delicate to quantify the consequences of each decision. We will see through paradoxes like the *Saint Petersburg paradox* that, even when the loss function seems obvious, for instance when errors can be expressed as monetary losses, the actual loss function can be quite different from its intuitive and linear approximation.

The complexity of determining the subjective loss function of the decision-maker often prompts the statistician to use classical (or *canonical*) losses, selected because of their simplicity and mathematical tractability. Such losses are also necessary for a theoretical treatment of the derivation of optimal procedures, when there is no practical motivation for the choice of a particular loss function. The term *classical* is related to their long history, dating back to Laplace (1773) for the absolute error loss (2.5.3) and Gauss (1810) for the quadratic loss (2.5.1), when *errors* in terms of performance of estimators or consequences of decisions were confused with *errors* in terms of the irreducible variability of random variables (variance). But this attribute should not be taken as a value statement, since an extensive use of these losses does not legitimize them any further. In fact, the recourse to such automatic (or generic) losses, although often justified in practice—it is still better to take a decision in a finite time using an approximate criterion rather that spending an infinite time to determine exactly the proper loss function—has generated many of the criticisms addressed to Decision Theory.

A fundamental basis of Bayesian Decision Theory is that statistical inference should start with the rigorous determination of three factors:

(1) the distribution family for the observations, $f(x|\theta)$;

(2) the prior distribution for the parameters, $\pi(\theta)$;

(3) the loss associated with the decisions, $\mathrm{L}(\theta, \delta)$;

the prior and the loss, and even sometimes the sampling distribution, being derived from partly subjective considerations. Classical decision-theoreticians omit the second point. The frequentist criticisms of the Bayesian paradigm often fail to take into account the problem of the construction of the loss function, even though this may be at least as complicated as the derivation of the prior distribution. In addition, to presuppose the existence of a loss function implies that some information about the problem at hand is available. This information could therefore be used more efficiently by building up a prior distribution. Actually, Lindley (1985) states that loss and prior are difficult to separate and should be analyzed simultaneously. We will see in Section 2.4 an example of the *duality* existing between these two factors. We also mention in Section 2.5.4 how classical losses could be replaced by more intrinsic losses (similar to the noninformative priors introduced in Chapter 3), when no information at all is available on the penalty associated with erroneous decisions or even with the parameterization of interest.

In some cases, it is possible to reduce the class of acceptable loss functions by *invariance* considerations, for example when the model is invariant under the action of a group of transformations. Such considerations apply as well to the choice of the prior distribution, as we will see in Chapter 9. It is also interesting to note that these invariance motivations are often used in other decision-theoretic approaches, where a drastic reduction of the class of inferential procedures is necessary to select a best solution.

Example 2.1.2 Consider the problem of estimating the mean θ of a normal vector, $x \sim \mathcal{N}_n(\theta, \Sigma)$, where Σ is a known diagonal matrix with diagonal elements σ_i^2 $(1 \leq i \leq n)$. In this case, $\mathcal{D} = \Theta = \mathbb{R}^p$, and δ stands for an evaluation of θ. If no additional information is available on the model, it seems logical to choose the loss function so that it weights equally the estimation of each component, i.e., to use a loss of the form

$$\sum_{i=1}^{n} \mathrm{L}\left(\frac{\delta_i - \theta_i}{\sigma_i}\right) ,$$

where L takes its minimum at 0. Indeed, for such losses, the components with larger variances do not strongly bias the selection of the resulting estimator. In other words, the components with a larger variance are not overly weighted when the estimation errors $(\delta_i - \theta_i)$ are normalized by σ_i. The usual choice of L is the quadratic loss $\mathrm{L}(t) = t^2$, i.e., the global estimation error is the sum of the squared componentwise errors. ‖

2.2 Existence of a utility function

The notion of utility (defined as the opposite of loss) is used not only in Statistics, but also in Economics and in other fields like Game Theory where it is necessary to *order* consequences of actions or decisions. *Consequences* (or *rewards*) are generic notions which summarize the set of outcomes resulting from the decision-maker's action. In the simplest cases, it may be the monetary profit or loss resulting from the decision. In an estimation setting, it may be a measure of distance between the evaluation and the true value of the parameter, as in Example 2.1.2. The axiomatic foundations of utility are due to Von Neumann and Morgenstern (1947) and led to numerous extensions, in particular in Game Theory. This approach is considered in a statistical framework by Wald (1950) and Ferguson (1967). Extensions and additional comments can be found in DeGroot (1970, Chapter 7) and recent references on utility theory are Fishburn (1988) and Machina (1982, 1987). See also Chamberlain (2000) for a connection with econometrics.

The general framework behind utility theory considers $\mathcal{R}$, space of *rewards*, which is assumed to be completely known. For instance, $\mathcal{R} = \mathbb{R}$. We also suppose that *it is possible to order the rewards*, i.e., that there exists a *total ordering*, denoted $\preceq$, on $\mathcal{R}$ such that, if r_1 and r_2 are in $\mathcal{R}$,

(1) $r_1 \preceq r_2$ or $r_2 \preceq r_1$; and

(2) if $r_1 \preceq r_2$ and $r_2 \preceq r_3$, then $r_1 \preceq r_3$.

These two properties seem to be minimal requirements in a decision-making setting. In particular, *transitivity* (2) is absolutely necessary to allow a comparison of decision procedures. Otherwise, we may end up with cycles such as $r_1 \preceq r_2 \preceq r_3 \preceq r_1$ and be at a loss about selecting the best reward among the three. Section 2.6 presents a criterion which is intransitive (and thus does not pertain to Decision Theory). We denote by $\prec$ and $\sim$ the *strict* order and *equivalence* relations derived from $\preceq$ respectively. Therefore, one and only one of the three following relations is satisfied by any pair (r_1, r_2) in $\mathcal{R}^2$

$$r_1 \prec r_2, \qquad r_2 \prec r_1, \qquad r_1 \sim r_2.$$

To proceed further in the construction of the utility function, it is necessary to extend the reward space from $\mathcal{R}$ to $\mathcal{P}$, the space of probability distributions on $\mathcal{R}$. This also allows the decision-maker to take into account partly randomized decisions; moreover, the extended reward space is convex.

Example 2.2.1 In most real-life settings, the rewards associated with an action are not exactly known when the decision is taken or, equivalently, some decisions involve a gambling step. For instance, in finance, the monetary revenue $r \in \mathcal{R} = \mathbb{R}$ derived from stock market shares is not guaranteed when the shareholder has to decide from which company she should buy shares. In this case, $\mathcal{D} = \{d_1, \ldots, d_n\}$, where d_k represents the action "buy the share from company k." At the time of the decision, the rewards

associated with the different shares are random dividends, only known by the end of the year. ‖

The order relation $\preceq$ is also assumed to be available on $\mathcal{P}$. For instance, when the rewards are monetary, the order relation on $\mathcal{P}$ can be derived by comparing the average yields associated with the distributions P. Therefore, it is possible to compare two distributions of probability on $\mathcal{R}$, P_1 and P_2. We thus assume that $\preceq$ satisfies the extensions of the two hypotheses (1) and (2) to $\mathcal{P}$:

(A_1) $P_1 \preceq P_2$ or $P_2 \preceq P_1$; and

(A_2) if $P_1 \preceq P_2$ and $P_2 \preceq P_3$, then $P_1 \preceq P_3$.

The order relation on $\mathcal{R}$ then appears as a special case of the order on $\mathcal{P}$ by considering the Dirac masses δ_r $(r \in \mathcal{R})$.

The existence of the order $\preceq$ on $\mathcal{P}$ relies on the assumption that there exists an optimal reward, therefore, that there exists at least a partial ordering on the consequences, even when they are random. This is obviously the case when there exists a function U on $\mathcal{R}$ associated with $\preceq$, such that $P_1 \preceq P_2$ is equivalent to

$$\mathbb{E}^{P_1}[U(r)] \leq \mathbb{E}^{P_2}[U(r)],$$

as in the above monetary example. This function U is called the *utility function*. We now present an axiomatic system on $\preceq$ that ensures the existence of the utility function.

For simplicity's sake, we only consider here the set of *bounded* distributions, $\mathcal{P}_B$, corresponding to the distributions with bounded support, for which there exist r_1 and r_2 such that

$$[r_1, r_2] = \{r : \ r_1 \preceq r \preceq r_2\} \quad \text{and} \quad P([r_1, r_2]) = 1.$$

For P_1, P_2 in $\mathcal{P}_B$, we define the *mixture* $P = \alpha P_1 + (1-\alpha)P_2$ as the distribution that generates a reward from P_1 with probability α and a reward from P_2 with probability $(1-\alpha)$. For instance, $\alpha r_1 + (1-\alpha)r_2$ is the distribution that gives the reward r_1 with probability α and the reward r_2 with probability $(1-\alpha)$. Two additional assumptions (or axioms) are necessary to derive the existence of a utility function on $\mathcal{R}$. First, there must be *conservation of the ordering under indifferent alternatives*:

(A_3) if $P_1 \preceq P_2$, $\alpha P_1 + (1-\alpha)P \preceq \alpha P_2 + (1-\alpha)P$ for every $P \in \mathcal{P}$.

For example, if the share buyers of Example 2.2.1 can compare two companies with dividend distributions P_1 and P_2, they should be able to keep a ranking of the two companies if there is a chance $(1-\alpha)$ that both dividends are replaced by state bounds with dividend distribution P. The order relation must also be *connected* (or *closed*):

(A_4) if $P_1 \preceq P_2 \preceq P_3$, there exist α and $\beta \in (0,1)$ such that
$\alpha P_1 + (1-\alpha)P_3 \preceq P_2 \preceq \beta P_1 + (1-\beta)P_3$.

The last assumption then implies the following result.

Lemma 2.2.2 *If r_1, r_2, and r are rewards in $\mathcal{R}$ with $r_1 \prec r_2$ and $r_1 \preceq r \preceq r_2$, there exists a unique v $(0 \le v \le 1)$ such that $r \sim vr_1 + (1-v)r_2$.*

Lemma 2.2.2 is actually the key to the derivation of the *utility function*, U, on $\mathcal{R}$. Indeed, given r_1 and r_2, two arbitrary rewards such that $r_2 \prec r_1$, we can define U in the following way. For every $r \in \mathcal{R}$, consider

(i) $U(r) = v$ if $r_2 \preceq r \preceq r_1$ and $r \sim vr_1 + (1-v)r_2$;

(ii) $U(r) = \frac{-v}{1-v}$ if $r \preceq r_2$ and $r_2 \sim vr_1 + (1-v)r$; and

(iii) $U(r) = \frac{1}{v}$ if $r_1 \preceq r$ and $r_1 \sim vr + (1-v)r_2$.

In particular, $U(r_1) = 1$ and $U(r_2) = 0$. Moreover, this function U preserves the order relation on $\mathcal{R}$ (see DeGroot (1970, p. 105) for a proof).

Lemma 2.2.3 *If r_1, r_2, and r_3 are three rewards in $\mathcal{R}$ such that $r_2 \sim \alpha r_1 + (1-\alpha)r_3$*

$$U(r_2) = \alpha U(r_1) + (1-\alpha)U(r_3).$$

Actually, the axioms (A_3) and (A_4) can be further reduced while Lemma 2.2.3 still holds. It is indeed sufficient that they are satisfied on $\mathcal{R}$ only. The extension of the definition of the utility function to $\mathcal{P}_B$ calls for an additional assumption. Given P such that $P([r_1, r_2]) = 1$, define

$$\alpha(r) = \frac{U(r) - U(r_1)}{U(r_2) - U(r_1)}$$

and

$$\beta = \int_{[r_1, r_2]} \alpha(r)\, dP(r).$$

Then the additional axiom

(A_5) $$P \sim \beta\delta_{r_2} + (1-\beta)\delta_{r_1}$$

implies that, if r is equivalent to $\alpha(r)r_1 + (1-\alpha(r))r_2$ for every $r \in [r_1, r_2]$, this equivalence must hold on average. In fact, notice that β is derived from the expected utility,

$$\beta = \frac{\mathbb{E}^P[U(r)] - U(r_1)}{U(r_2) - U(r_1)},$$

and this assumption provides a definition of U on $\mathcal{P}_B$. As in Lemma 2.2.3 where U is restricted to $\mathcal{R}$, and as shown by the following result, axiom (A_5) indicates that U provides a *linearization* (or a linear parameterization) of the order relation $\preceq$ on $\mathcal{P}_B$. Although slightly tautological—since it involves in its formulation the utility function we are trying to derive—, (A_5) indeed leads to the following extension of Lemma 2.2.3 to $\mathcal{P}_B$.

Theorem 2.2.4 *Consider P_1 and P_2 in $\mathcal{P}_B$. Then,*

$$P_1 \preceq P_2$$

if and only if

$$\mathbb{E}^{P_1}[U(r)] \le \mathbb{E}^{P_2}[U(r)].$$

Moreover, if U^* is another utility function satisfying the above equivalence relation, there exist $a > 0$ and b such that

$$U^*(r) = aU(r) + b.$$

Proof. Consider r_1 and r_2 such that

$$P_1([r_1, r_2]) = P_2([r_1, r_2]) = 1$$

(with $r_1 \prec r_2$). Since

$$P_1 \sim \frac{\mathbb{E}^{P_1}[U(r)] - U(r_1)}{U(r_2) - U(r_1)} r_2 + \frac{U(r_2) - \mathbb{E}^{P_1}[U(r)]}{U(r_2) - U(r_1)} r_1$$

and

$$P_2 \sim \frac{\mathbb{E}^{P_2}[U(r)] - U(r_1)}{U(r_2) - U(r_1)} r_2 + \frac{U(r_2) - \mathbb{E}^{P_2}[U(r)]}{U(r_2) - U(r_1)} r_1,$$

$P_1 \preceq P_2$ is truly equivalent to

$$\frac{\mathbb{E}^{P_1}[U(r)] - U(r_1)}{U(r_2) - U(r_1)} \le \frac{\mathbb{E}^{P_2}[U(r)] - U(r_1)}{U(r_2) - U(r_1)},$$

i.e., $\mathbb{E}^{P_1}[U(r)] \le \mathbb{E}^{P_2}[U(r)]$. Moreover, for any other utility function U^*, there exist a and b such that $U^*(r_1) = aU(r_1) + b$, $U^*(r_2) = aU(r_2) + b$. The extension of this relation to every $r \in \mathcal{R}$ follows from Lemma 2.2.3. □□

Notice that the above derivation does not involve any restriction on the function U. Therefore, it does not need to be bounded, although this condition is often mentioned in textbooks. It can be argued that this generality is artificial and formal, since subjective utility functions are always bounded. For instance, when considering monetary rewards, there is a psychological threshold, say \$100,000,000, above which (most) individuals have an almost constant utility function.

However, this upper bound varies from individual to individual, and even more so from individuals to companies or states. It is also important to incorporate unacceptable rewards, although the assumption (A_4) prevents rewards with utility equal to $-\infty$. (This restriction implies that the death of a patient in a pharmaceutical study or a major accident in a nuclear plant have a finite utility.) Moreover, most theoretical losses are not bounded. A counterpart of this generality is that the above results have only been established for $\mathcal{P}_B$. Actually, they can be extended to $\mathcal{P}_\mathcal{E}$, the set of distributions P in $\mathcal{P}$ such that $\mathbb{E}^P[U(r)]$ is finite, under the assumption that (A_1)–(A_5) and two additional hypotheses are satisfied for $\mathcal{P}_\mathcal{E}$ (see Exercise 2.3).

Theorem 2.2.5 *Consider P and Q, two distributions in $\mathcal{P}_\mathcal{E}$. Then, $P \preceq Q$ if and only if*

$$\mathbb{E}^P[U(r)] \le \mathbb{E}^Q[U(r)].$$

Of course, Theorem 2.2.5 fails to deal with infinite utility distributions. If such distributions exist, they must be compared between themselves and a separate utility function constructed on this restricted class, since they are in a sense the only distributions of interest. However, the loss functions considered in the sequel are bounded from below, usually by 0. Therefore, the corresponding utility functions, opposites of the loss functions, are always bounded from above and infinite reward paradoxes can be avoided. (Rubin (1984) and Fishburn (1987) provide reduced axiomatic systems ensuring the existence of a utility function.)

Many criticisms have been addressed on theoretical or psychological grounds against the notion of *rationality of decision-makers* and the associated axioms (A_1)–(A_4). First, it seems illusory to assume that individuals can compare all rewards, that is, that they can provide a total ordering of $\mathcal{P}$ (or even of $\mathcal{R}$) because their discriminating abilities are necessarily limited, especially about contiguous or extreme alternatives. The *transitivity* assumption is also too strong, since examples in sports or politics show that real-life orderings of preferences often lead to nontransitivity, as illustrated by *Condorcet and Simpson* paradoxes (see Casella and Wells (1993) and Exercises 1.9 and 2.2). More fundamentally, the assumption that the ordering can be extended from $\mathcal{R}$ to $\mathcal{P}$ has been strongly attacked because it implies that a social ordering can be derived from a set of individual orderings and this is not possible in general (see Arrow (1951) or Blyth (1993)). However, while recognizing this fact, Rubin (1987) notes that this impossibility just implies that utility and prior are not separable, not that an optimal (Bayesian) decision cannot be obtained, and he gives a restricted set of axioms pertaining to this purpose. In general, the criticisms above are obviously valuable, but cannot stand against the absolute need of an axiomatic framework validating decision-making under uncertainty. As already mentioned in Chapter 1, statistical modeling *is and must be* reductive; although necessarily missing part of the complexity of the world, the simplified representation it gives of this world allows statisticians and others to reach decisions. Decision Theory thus describes an idealized setting, under an ultimate rationality real decision-makers fail to attain, but aim at nonetheless[2].

From a more practical point of view, the above derivation of the utility function can be criticized as being unrealistic. Berger (1985a) provides a few examples based on DeGroot (1970), deriving the utility function from successive partitions of the reward space (see also Raiffa and Schlaifer (1961)). However, if $\mathcal{R}$ is large (e.g., noncountable), U cannot be evaluated for each reward r, even though the linearity exhibited by Lemma 2.2.3 allows for approximations when $\mathcal{R} \subset \mathbb{R}$. In a multidimensional setting, linear approximations are no longer possible unless one uses a linear combination of

[2] To borrow from Smith (1984), to criticize the idealized structures of Decision Theory because of human limitations is somehow akin to attacking integration because some integrals can only be solved numerically.

componentwise utilities, i.e.,

$$U(r_1, r_2, \ldots, r_n) = \sum_{i=1}^{n} \alpha_i U_i(r_i)$$

(see Raiffa (1968), Keeney and Raifa (1976) and Smith (1988) for a discussion). In general, practical utility functions will thus only approximate the true utility functions.

Even when the reward is purely monetary there is a necessity of rigorous determination of the utility function because U may be far from linear, especially for large rewards. This means that a gain of \$3000 with probability 1/2 may not be equivalent to earning \$1500 with certainty. To solve this paradox, Laplace (1795) introduced the notion of *moral expectation*, derived from the relative value of an increase of wealth, *"absolute value divided by the total wealth of the involved person."* Laplace deduces that the moral expectation *"coincides with the mathematical expectation when the wealth becomes infinite compared with the variations due to uncertainty,"* meaning the utility is indeed linear only around 0. Otherwise, *risk aversion* attitudes slow down the utility curve, which is typically concave for large values of rewards and bounded above. (Persons with a convex utility function are called *risk lovers* because they prefer a random gain to the expectation of this gain. Notice that this attitude is quite understandable in a neighborhood of 0.) To construct the money utility function is obviously more cumbersome than to use a linear utility, but this derivation gives a more accurate representation of reality and can even prevent paradoxes such as the following one.

Example 2.2.6 (Saint Petersburg Paradox) Consider a game where a coin is thrown until a *head* appears. When this event occurs at the nth throw, the player gain is 3^n, leading to an average gain of

$$\sum_{n=1}^{+\infty} 3^n \frac{1}{2^n} = +\infty.$$

Every player should then be ready to pay an arbitrarily high entrance fee to play this game, even though there is less than a 0.05 chance to go beyond the fifth throw! This modeling does not take into account that the fortune of a player is necessarily bounded and that he or she can only play a limited number of games. A solution to this paradox is to substitute for the linear utility function a bounded utility function, such as

$$U(r) = \frac{r}{\delta + r} \qquad (\delta > 0,\ r > -\delta),$$

and $U(r) = -\infty$ otherwise. This construction is quite similar to Laplace's moral expectation. An acceptable entrance fee e will then be such that the expected utility of the game is larger than the utility of doing nothing, i.e.,

$$\mathbb{E}[U(r - e)] \geq U(0) = 0.$$

Consider now a modification to the game such that the player can leave the game at any time n and take the gain 3^n if a head has not yet appeared. The average gain at time n is then

$$\frac{3^n}{\delta + 3^n} 2^{-n},$$

which can provide an optimal leaving time n_0, depending on the utility parameter δ, which in its turn somehow characterizes the *risk aversion* of the player (see Smith (1988) for a more thorough description). For instance, δ may represent the fortune of the player, since $U(\tau)$ goes to $-\infty$ when τ goes to $-\delta$. The particular choice of U can obviously be criticized, but a more accurate representation of the utility function requires a detailed analysis of the motivations of the player (see also Exercise 2.9). ‖

See also Bernardo and Smith (1994) for a detailed analysis of the foundations of Utility Theory, with in particular a description of *decision trees*.

2.3 Utility and loss

Let us switch back to a purely statistical setting. From a decision-theoretic point of view, the statistical model now involves three spaces: $\mathcal{X}$, *observation* space, Θ, *parameter* space, and $\mathcal{D}$, *decision* space (or *action* space). Statistical inference then consists of taking a decision $d \in \mathcal{D}$ related to the parameter $\theta \in \Theta$ based on the observation $x \in \mathcal{X}$, x and θ being related by the distribution $f(x|\theta)$. In most cases, the decision d will be to evaluate (or *estimate*) a function of θ, $h(\theta)$, as accurately as possible. Decision Theory assumes in addition that each action d can be evaluated (i.e., that its accuracy can be quantified) and leads to a reward r, with utility $U(r)$ (which exists under the assumption of rationality of the decision-makers). From now on, this utility is written as $U(\theta, d)$ to stress that it only depends on these two factors. When other random factors r are involved in U, we take $U(\theta, d) = \mathbb{E}_{\theta,d}[U(r)]$. Therefore, $U(\theta, d)$ can be seen as a measure of proximity between the proposed estimate d and the true value $h(\theta)$.

Once the utility function has been constructed (or approximated), we derive the corresponding *loss* function

$$\mathrm{L}(\theta, d) = -U(\theta, d).$$

In general, the loss function is supposed to be nonnegative, which implies that $U(\theta, d) \leq 0$, and therefore that there is no decision with infinite utility. The existence of a lower bound on L can be criticized as being too stringent, but it does avoid paradoxes such as those mentioned above. It can also be argued that, from a statistical point of view, the loss function L indeed represents the *loss* (or error) owing to a bad evaluation of the function of θ of interest, and therefore that even the best evaluation of this function, i.e., when θ is known, can induce at best a null loss. Otherwise, there would be

a lack of continuity of the loss function in $d = \theta$ which could even prevent the choice of a decision procedure.

Obviously, except for the most trivial settings, it is generally impossible to uniformly minimize (in d) the loss function $\mathrm{L}(\theta, d)$ when θ is unknown. In order to derive an effective comparison criterion from the loss function, the *frequentist* approach proposes to consider instead the average loss (or *frequentist risk*)

$$\begin{aligned} R(\theta, \delta) &= \mathbb{E}_\theta[\mathrm{L}(\theta, \delta(x))] \\ &= \int_{\mathcal{X}} \mathrm{L}(\theta, \delta(x)) f(x|\theta)\, dx, \end{aligned}$$

where $\delta(x)$ is the decision rule, i.e., the allocation of a decision to each outcome $x \sim f(x|\theta)$ from the random experiment. The function δ, from $\mathcal{X}$ in $\mathcal{D}$, is usually called *estimator* (while the *value* $\delta(x)$ is called *estimate* of θ). When there is no risk of confusion, we also denote the set of estimators by $\mathcal{D}$.

The frequentist paradigm relies on this criterion to compare estimators and, if possible, to select the best estimator, the reasoning being that estimators are evaluated on their long-run performance for all possible values of the parameter θ. Notice, however, that there are several difficulties associated with this approach.

(1) The error (loss) is averaged over the different values of x proportionally to the density $f(x|\theta)$. Therefore, it seems that the observation x is not taken into account any further. The risk criterion evaluates procedures on their long-run performance and not directly for the given observation, x. Such an evaluation may be satisfactory for the statistician, but it is not so appealing for a client, who wants optimal results for her data x, not that of another's!

(2) The frequentist analysis of the decision problem implicitly assumes that this problem will be met again and again, for the frequency evaluation to make sense. Indeed, $R(\theta, \delta)$ is approximately the average loss over i.i.d. repetitions of the same experiment, according to the Law of Large Numbers. However, on both philosophical and practical grounds, there is a lot of controversy over the very notion of *repeatability of experiments* (see Jeffreys (1961)). For one thing, if new observations come to the statistician, she should make use of them, and this could modify the way the experiment is conducted, as in, for instance, medical trials.

(3) For a procedure δ, the risk $R(\theta, \delta)$ is a *function* of the parameter θ. Therefore, the frequentist approach does not induce a total ordering on the set of procedures. It is generally impossible to compare decision procedures with this criterion, since two crossing risk functions prevent comparison between the corresponding estimators. At best, one may hope for a procedure δ_0 that uniformly minimizes $R(\theta, \delta)$, but such cases rarely occur unless the space of decision procedures is restricted.

Best procedures can only be obtained by restricting rather artificially the set of authorized procedures.

Example 2.3.1 Consider x_1 and x_2, two observations from

$$P_\theta(x = \theta - 1) = P_\theta(x = \theta + 1) = 0.5, \qquad \theta \in \mathbb{R}.$$

The parameter of interest is θ (i.e., $\mathcal{D} = \Theta$) and it is estimated by estimators δ under the loss

$$\mathrm{L}(\theta, \delta) = 1 - \mathbb{I}_\theta(\delta),$$

often called $0-1$ *loss*, which penalizes errors of estimation, whatever their magnitude, by 1. Considering the particular estimator

$$\delta_0(x_1, x_2) = \frac{x_1 + x_2}{2},$$

its risk function is

$$\begin{aligned} R(\theta, \delta_0) &= 1 - P_\theta(\delta_0(x_1, x_2) = \theta) \\ &= 1 - P_\theta(x_1 \neq x_2) = 0.5. \end{aligned}$$

This computation shows that the estimator δ_0 is correct half of the time. Actually, this estimator is always correct when $x_1 \neq x_2$, and always wrong otherwise. Now, the estimator $\delta_1(x_1, x_2) = x_1 + 1$ also has a risk function equal to 0.5, as does $\delta_2(x_1, x_2) = x_2 - 1$. Therefore, δ_0, δ_1 and δ_2 cannot be ranked under the $0-1$ loss. ‖

On the contrary, the Bayesian approach to Decision Theory integrates on the space Θ since θ is unknown, instead of integrating on the space $\mathcal{X}$ as x is known. It relies on the *posterior expected loss*

$$\begin{aligned} \varrho(\pi, d|x) &= \mathbb{E}^\pi[L(\theta, d)|x] \\ &= \int_\Theta \mathrm{L}(\theta, d)\pi(\theta|x)\, d\theta, \end{aligned}$$

which averages the error (i.e., the loss) according to the posterior distribution of the parameter θ, *conditionally on the observed value* x. Given x, the average error resulting from decision d is actually $\varrho(\pi, d|x)$. The posterior expected loss is thus a function of x but this dependence is not troublesome, as opposed to the frequentist dependence of the risk on the parameter because x, contrary to θ, is known.

Given a prior distribution π, it is also possible to define the *integrated risk*, which is the frequentist risk averaged over the values of θ according to their prior distribution

$$\begin{aligned} r(\pi, \delta) &= \mathbb{E}^\pi[R(\theta, \delta)] \\ &= \int_\Theta \int_\mathcal{X} \mathrm{L}(\theta, \delta(x))\, f(x|\theta)\, dx\ \pi(\theta)\, d\theta. \end{aligned}$$

One particular interest of this second concept is that it associates a real number with every estimator, not a function of θ. It therefore induces a

total ordering on the set of estimators, i.e., allows for the direct comparison of estimators. This implies that, while taking into account the prior information through the prior distribution, the Bayesian approach is sufficiently *reductive* (in a positive sense) to reach an effective decision. Moreover, the above two notions are equivalent in that they lead to the same decision.

Theorem 2.3.2 *An estimator minimizing the integrated risk $r(\pi, \delta)$ can be obtained by selecting, for every $x \in \mathcal{X}$, the value $\delta(x)$ which minimizes the posterior expected loss, $\varrho(\pi, \delta|x)$, since*

$$r(\pi, \delta) = \int_\mathcal{X} \varrho(\pi, \delta(x)|x) m(x)\, dx. \tag{2.3.1}$$

Proof. Equality (2.3.1) follows directly from Fubini's Theorem since, as $\mathrm{L}(\theta, \delta) \geq 0$,

$$\begin{aligned} r(\pi, \delta) &= \int_\Theta \int_\mathcal{X} \mathrm{L}(\theta, \delta(x)) f(x|\theta)\, dx\ \pi(\theta)\, d\theta \\ &= \int_\mathcal{X} \int_\Theta \mathrm{L}(\theta, \delta(x)) f(x|\theta)\pi(\theta)\, d\theta\, dx \\ &= \int_\mathcal{X} \int_\Theta \mathrm{L}(\theta, \delta(x))\pi(\theta|x)\, d\theta\ m(x)\, dx\ . \end{aligned}$$

□□

This result leads to the following definition of a Bayes estimator.

Definition 2.3.3 *A Bayes estimator associated with a prior distribution π and a loss function* L *is any estimator δ^π which minimizes $r(\pi, \delta)$. For every $x \in \mathcal{X}$, it is given by $\delta^\pi(x)$, argument of* $\min_d \varrho(\pi, d|x)$. *The value $r(\pi) = r(\pi, \delta^\pi)$ is then called the Bayes risk.*

Theorem 2.3.2 thus provides a constructive tool for the determination of the Bayes estimators. Notice that, from a strictly Bayesian point of view, only the posterior expected loss $\varrho(\pi, \delta|x)$ is important, as the Bayesian paradigm is based on the conditional approach. To average over all possible values of x, when we know the observed value of x, seems to be a waste of information. Nonetheless, the equivalence exhibited in Theorem 2.3.2 is important because, on one hand, it shows that the conditional approach is not necessarily as dangerous as frequentists may depict it. This is so because, while the Bayesian approach works conditional upon the actual observation x, it also incorporates the probabilistic properties of the distribution of the observation, $f(x|\theta)$. On the other hand, this equivalence provides a connection between the classical results of Game Theory (see Section 2.4) and the axiomatic Bayesian approach, based on the posterior distribution. It also explains why Bayes estimators play an important role in frequentist optimality criteria.

The result above is valid for proper and improper priors, *as long as the Bayes risk $r(\pi)$ is finite*. Otherwise, the notion of a (decision-theoretic)

Bayes estimator is weakened: we then define a *generalized Bayes estimator* as the minimizer, for every x, of the posterior expected loss. In terms of frequentist optimality, we will see that the division between proper and improper priors is much less important than the division between regular and generalized Bayes estimators, since the formers are admissible. Notice that, for strictly convex losses, the Bayes estimators are unique.

We conclude this section with an example of construction of a loss function in an expert calibration framework. References on this topic are DeGroot and Fienberg (1983), Murphy and Winkler (1984), Bayarri and DeGroot (1988) and Schervish (1989). Smith (1988) also shows how forecaster evaluation can help improve the assessment of prior probabilities. See also Note 2.8.1 for an illustration in imaging.

Example 2.3.4 Meteorological forecasts are often given as probability statements such as *"the probability of rain for tomorrow is 0.4."* Such forecasts being quantified, it is of interest to evaluate weather forecasters through a loss function (for their employers as well as users).

For a given forecaster, let N be the number of different percentages predicted at least once in a year and let p_i $(1 \le i \le N)$ be the corresponding percentages. For instance, we may have $N = 5$ and

$$p_1 = 0, \quad p_2 = 0.45, \quad p_3 = 0.7, \quad p_4 = 0.9, \quad \text{and} \quad p_5 = 0.95.$$

In this case, the parameters θ_i are actually observed, i.e.,

$$\theta_i = \frac{\text{number of rainy days when } p_i \text{ is forecasted}}{\text{number of days when } p_i \text{ is forecasted}}$$

(more exactly, this ratio is a good approximation of θ_i).

If q_i denotes the proportion of days where p_i is forecasted, a possible loss function for the forecasters is

$$\mathrm{L}(\theta, p) = \sum_{i=1}^{N} q_i (p_i - \theta_i)^2 + \sum_{i=1}^{N} q_i \log(q_i).$$

For a given set of θ_i's $(1 \le i \le N)$, the best forecaster is the perfectly calibrated forecaster, i.e., the one who satisfies $p_i = \theta_i$ $(1 \le i \le N)$. Moreover, among these perfect forecasters, the best one is the most well balanced, satisfying $q_i = 1/N$ $(1 \le i \le N)$, i.e., the more daring forecaster, as opposed to a forecaster which would always give the same forecast, p_{i_0}, because of the *entropy* term, $\sum_i q_i \log(q_i)$. However, the distance $(p_i - \theta_i)^2$ could be replaced by any other function taking its minimum at $p_i = \theta_i$ (see Exercises 2.12 and 2.14). The weight q_i in the first sum is also used to calibrate more properly forecasters, in order to prevent overpenalization of rare forecasts.

This loss has been constructed with a bias in favor of forecasters with large N, since the *entropy* $\log(N)$ increases with N. However, a better performance for a larger N requires that p_i is (almost) equal to θ_i and q_i is close to $1/N$. ‖

2.4 Two optimalities: minimaxity and admissibility

This section deals with the two fundamental notions of frequentist Decision Theory, introduced by Wald (1950) and Neyman and Pearson (1933a,b). As mentioned above, and contrary to the Bayesian approach, the frequentist paradigm is not reductive enough to lead to a single optimal estimator. While we are mainly concerned in this book with the Bayesian aspects of Decision Theory, it is still necessary to study these frequentist notions in detail because they show that Bayes estimators are often optimal for the frequentist concepts of optimality, therefore should still be considered even when prior information is ignored. In other words, one can reject the Bayesian paradigm and ignore the meaning of the prior distribution and still obtain good estimators from a frequentist point of view when using this prior distribution. Therefore, in this technical sense, frequentists should also take into account the Bayesian approach, since it provides a *tool* for the derivation of optimal estimators (see Brown (1971, 2000), Strawderman (1974), Berger (1985a), or Berger and Robert (1990) for examples). Moreover, these properties can be helpful in the selection of a prior distribution, when prior information is not precise enough to lead to a single prior distribution (see Chapter 3).

2.4.1 Randomized estimators

Similar to the study of the utility function, where we extended the reward space from $\mathcal{R}$ to $\mathcal{P}$, we need to extend the decision space to the set of *randomized estimators*, taking values in $\mathcal{D}^*$, space of the probability distributions on $\mathcal{D}$. To use a randomized estimator δ^* means that the action is generated according to the distribution with probability density $\delta^*(x, .)$, once the observation x has been collected. The loss of a randomized estimator δ^* is then defined as the average loss

$$\mathrm{L}(\theta, \delta^*(x)) = \int_{\mathcal{D}} \mathrm{L}(\theta, a) \delta^*(x, a)\, da.$$

This extension is necessary to deal with minimaxity and admissibility. Obviously, such estimators are not to be used, if only because they contradict the Likelihood Principle, giving several possible answers for the same value of x (and thus of $\ell(\theta|x)$). Moreover, it seems quite paradoxical to add noise to a phenomenon in order to take a decision under uncertainty!

Example 2.4.1 (Example 2.3.1 continued) Consider the randomized estimator

$$\delta^*(x_1, x_2)(t) = \begin{cases} \mathbb{I}_{(x_1+x_2)/2}(t) & \text{if } x_1 \neq x_2, \\ [\mathbb{I}_{(x_1-1)}(t) + \mathbb{I}_{(x_1+1)}(t)]/2 & \text{otherwise,} \end{cases}$$

where $\mathbb{I}_v$ denotes the Dirac mass at v. Actually, if $x_1 = x_2$, the two values $\theta_1 = x_1 - 1$ and $\theta_2 = x_1 + 1$ have the same likelihood. Compared with δ_0 which never estimates θ correctly if $x_1 = x_2$, δ^* is exact with probability 1/2. However, when δ^* misses θ, it is farther away from θ than δ_0. The choice of the estimator then depends on the loss function, i.e., the way the distance between the estimator and θ (or the error) is measured. ‖

Randomized estimators are nonetheless necessary from a frequentist point of view, for instance, for the frequentist theory of tests, as they provide access to confidence levels otherwise unattainable (see Chapter 5). The set $\mathcal{D}^*$ thus appears as a completion of $\mathcal{D}$. However, this modification of the decision space does not modify the Bayesian answers, as shown by the following result (where $\mathcal{D}^*$ also denotes the set of functions taking values in $\mathcal{D}^*$).

Theorem 2.4.2 *For every prior distribution π on Θ, the Bayes risk on the set of randomized estimators is the same as the Bayes risk on the set of nonrandomized estimators, i.e.,*

$$\inf_{\delta\in\mathcal{D}} r(\pi,\delta) = \inf_{\delta^*\in\mathcal{D}^*} r(\pi,\delta^*) = r(\pi).$$

Proof. For every $x \in \mathcal{X}$ and every $\delta^* \in \mathcal{D}^*$, we have

$$\begin{aligned}
&\int_\Theta \int_{\mathcal{D}} L(\theta,a)\delta^*(x,a)da\,\pi(\theta|x)d\theta \\
&= \int_{\mathcal{D}} \int_\Theta L(\theta,a)\pi(\theta|x)d\theta\,\delta^*(x,a)da \\
&\geq \int_{\mathcal{D}} \inf_a \left\{ \int_\Theta L(\theta,a)\pi(\theta|x)d\theta \right\} \delta^*(x,a)da \\
&= \varrho(\pi,\delta^\pi|x).
\end{aligned}$$

□□

This result thus holds even when the Bayes risk $r(\pi)$ is infinite. The proof relies on the fact that a randomized procedure averages the risks of nonrandomized estimators and thus cannot improve on them. However, the fact that randomized procedures are not relevant does not hold for the frequentist risk unless some conditions, such as convexity, are imposed on the loss function.

2.4.2 Minimaxity

The minimax criterion we introduce now appears as an insurance against the worst case because it aims at minimizing the expected loss in the least favorable case. It also represents a frequentist effort to skip the Bayesian paradigm while producing a (weak) total ordering on $\mathcal{D}^*$.

Definition 2.4.3 *The minimax risk associated with a loss function* L *is the value*

$$\bar{R} = \inf_{\delta\in\mathcal{D}^*} \sup_\theta R(\theta,\delta) = \inf_{\delta\in\mathcal{D}^*} \sup_\theta \mathbb{E}_\theta[L(\theta,\delta(x))],$$

and a minimax estimator is any (possibly randomized) estimator δ_0 such that

$$\sup_\theta R(\theta,\delta_0) = \bar{R}.$$

This notion is validated by Game Theory, where two adversaries (here, "the statistician" and "Nature") are competing. Once the statistician has selected a procedure, Nature selects the state of nature (i.e., the parameter) that maximizes the loss of the statistician. (We will see below that this choice is usually equivalent to the choice of a prior distribution π. Therefore, the Bayesian approach does not really fit in that conflicting framework, since the prior distribution is also supposed to be known.) In general, it seems unfortunate to resort to such an antagonistic perspective in a statistical analysis. Indeed, to perceive Nature (or reality) as an enemy involves a bias toward the worst cases and prevents the statistician from using the available information (for an analysis and a defense of minimaxity, see Brown (1993) and Strawderman (2000)).

The notion of minimaxity provides a good illustration of the conservative aspects of the frequentist paradigm. Since this approach refuses to make any assumption on the parameter θ, it has to consider the worst cases as equally likely, and thus needs to focus on the maximal risk. In fact, from a Bayesian point of view, it is often equivalent to take a prior concentrated on these worst cases (see Section 2.4.3). In most settings, this point of view is thus too conservative because some values of the parameter are less likely than others.

Example 2.4.4 The first oil-drilling platforms in the North Sea were designed according to a minimax principle. In fact, they were supposed to resist the conjugate action of the worst gale and the worst storm ever observed, at the minimal record temperature. This strategy obviously gives a comfortable margin of safety, but is quite costly. For more recent platforms, engineers have taken into account the distribution of these weather phenomena in order to reduce the production cost. ‖

Example 2.4.5 A waiting queue at a red light is usually correctly represented by a Poisson distribution. The number of cars arriving during the observation time, N, is thus distributed according to $\mathcal{P}(\lambda)$, with the mean parameter λ to be estimated. Obviously, the values of λ above a given limit are quite unlikely. For instance, if λ_0 is the number of cars in the whole city, the average number of cars waiting at a given traffic light will not exceed λ_0. However, it may happen that some estimators are not minimax because their risk are above $\bar{R}$ for the largest values of λ. ‖

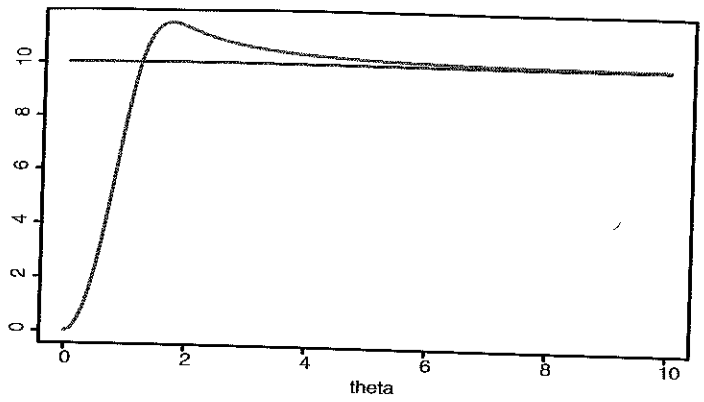

Figure 2.4.1. *Comparison of the risks of the estimators δ_1 and δ_2.*

The example above does not directly criticize the minimax principle but rather argues for the fact that some residual information is attached to most problems and that it should be used, even marginally. In a similar manner, Example 2.4.6 exhibits two estimators, δ_1 and δ_2, such that δ_1 has a constant minimax risk $\bar{R}$ and δ_2 has a risk which can be as low as $\bar{R}/10$ but goes slightly above $\bar{R}$ for the largest values of the parameter (see Figure 2.4.2). Therefore, according to the minimax principle, δ_1 should be preferred to δ_2, although the values of θ for which δ_1 dominates δ_2 are the most unlikely (see Exercise 2.28 for another striking example).

Example 2.4.6 For reasons explained in Note 2.8.2, we consider the following estimator

$$\delta_2(x) = \begin{cases} \left(1 - \dfrac{2p-1}{||x||^2}\right) x & \text{if } ||x||^2 \geq 2p-1, \\ 0 & \text{otherwise,} \end{cases}$$

to estimate θ when $x \sim \mathcal{N}_p(\theta, I_p)$. This estimator, called the *positive-part James–Stein estimator*, is evaluated under *quadratic loss*,

$$\mathrm{L}(\theta, d) = ||\theta - d||^2.$$

Figure 2.4.2 gives a comparison of the respective risks of δ_2 and $\delta_1(x) = x$, maximum likelihood estimator, for $p = 10$. This figure shows that δ_2 cannot be minimax, since the maximum risk of δ_2 is above the (constant) risk of δ_1, that is, $R(\theta, \delta_2) = \mathbb{E}_\theta[||\theta - \delta_2(x)||^2] = p$. (We show in Section 2.4.3 that δ_1 is actually minimax in this case.) But the estimator δ_2 is definitely superior on the most interesting part of the parameter space, the additional loss being in perspective quite negligible. ‖

Table 2.4.1. *Utility function $U(\theta_i, a_j)$.*

	a_1	a_2
θ_1	−4	− 10
θ_2	8	30

The opposition between minimax and Bayesian analyses is illustrated by the following example, which borrows from Game Theory (since there is no observation or statistical model).

Example 2.4.7 Two persons, A and B, suspected of being accomplices in a robbery, have been apprehended and placed in separate cells. Both suspects are questioned and enticed to confess the burglary. Although they cannot be convicted unless one of them talks, the incentive is that the first person to cooperate will get a reduced sentence. Table 2.4.7 provides the rewards as perceived by A (in years of freedom), where a_1 (resp. θ_1) represents the fact that A (resp. B) talks. The two suspects have an optimal gain if they both remain silent. However, from A's point of view, the optimal strategy is to be the first one to talk, i.e., a_1, since $\max_\theta R(a_1, \theta) = 4$ and $\max_\theta R(a_2, \theta) = 10$. Therefore, both burglars will end up in jail!

On the contrary, if π is the (subjective) probability assigned by A to the event *"B talks,"* i.e., to θ_1, the Bayes risk of a_1 is

$$r(\pi, a_1) = \mathbb{E}^\pi[-U(\theta, a_1)] = 4\pi - 8(1-\pi) = 12\pi - 8$$

and, for a_2,

$$r(\pi, a_2) = \mathbb{E}^\pi[-U(\theta, a_2)] = 10\pi - 30(1-\pi) = 40\pi - 30\,.$$

It is straightforward to check that, for $\pi \leq 11/14$, $r(\pi, a_2)$ is smaller than $r(\pi, a_1)$. Therefore, unless A is convinced that B will talk, it is better for A to keep silent. ‖

2.4.3 Existence of minimax rules and maximin strategy

An important difficulty related with minimaxity is that a minimax estimator does not necessarily exist. Ferguson (1967) and Berger (1985a, Chapter 5) give sufficient conditions. In particular, there exists a minimax strategy when Θ is finite and the loss function is continuous. More generally, Brown (1976) (see also Le Cam (1986) and Strasser (1985)) considers the decision space $\mathcal{D}$ as embedded in another space so that the set of risk functions on $\mathcal{D}$ is compact in this larger space. From this perspective and under additional assumptions, it is then possible to derive minimax estimators when the loss is continuous. However, these extensions involve topological techniques too advanced to be considered in this book. Therefore, we only give the following result (see Blackwell and Girshick (1954) for a proof).

Theorem 2.4.8 *If $\mathcal{D} \subset \mathbb{R}^k$ is a convex compact set and if $L(\theta, d)$ is continuous and convex as a function of d for every $\theta \in \Theta$, there exists a nonrandomized minimax estimator.*

The restriction to nonrandomized estimators when the loss is convex follows from *Jensen's inequality*, since

$$L(\theta, \delta^*) = \mathbb{E}^{\delta^*}[L(\theta, \delta)] \geq L(\theta, \mathbb{E}^{\delta^*}(\delta)).$$

This result is a special case of the *Rao–Blackwell Theorem* (see Lehmann and Casella (1998, p. 47)).

Example 2.4.9 (Example 2.4.1 continued) The randomized estimator δ^* is uniformly dominated for every convex loss by the nonrandomized estimator $\mathbb{E}^{\delta^*}[\delta^*(x_1, x_2)]$, i.e.,

$$\tilde{\delta}(x_1, x_2) = \begin{cases} \frac{1}{2}(x_1 + x_2) & \text{if } x_1 \neq x_2, \\ \frac{1}{2}(x_1 - 1) + \frac{1}{2}(x_1 + 1) = x_1 & \text{otherwise,} \end{cases}$$

which is actually identical to the estimator δ_0 considered originally. Notice that this is not true for the $0-1$ loss where δ^* dominates $\tilde{\delta}$. ‖

The following result points out the connection between the Bayesian approach and the minimax principle. (The proof is straightforward and thus omitted.)

Lemma 2.4.10 *The Bayes risks are always smaller than the minimax risk, i.e.,*

$$\underline{R} = \sup_{\pi} r(\pi) = \sup_{\pi} \inf_{\delta \in \mathcal{D}} r(\pi, \delta) \leq \bar{R} = \inf_{\delta \in \mathcal{D}^*} \sup_{\theta} R(\theta, \delta).$$

The first value is called *maximin risk* and a distribution π^* such that $r(\pi^*) = \underline{R}$ is called a *least favorable distribution*, when such distributions exist. In general, the upper bound $r(\pi^*)$ is rather attained by an improper distribution, which can be expressed as a limit of proper prior distributions π_n, but this phenomenon does not necessarily deter from the derivation of minimax estimators (see Lemma 2.4.15). When they exist, least favorable distributions are those with the largest Bayes risk, thus the less interesting distributions in terms of loss performances if they are not suggested by the available prior information. The above result is quite logical, in the sense that prior information can only improve the estimation error, even in the worst case.

A particularly interesting case corresponds to the following definition.

Definition 2.4.11 *The estimation problem is said to have a value when $\underline{R} = \bar{R}$, i.e., when*

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

When the problem has a value, some minimax estimators are the Bayes estimators for the least favorable distributions. However, they may be randomized, as illustrated by the following example. Therefore, the minimax principle does not always lead to acceptable estimators.

Example 2.4.12 Consider[3] a Bernoulli observation, $x \sim \mathcal{B}e(\theta)$ with $\theta \in \{0.1, 0.5\}$. Four nonrandomized estimators are available,

$$\begin{aligned} \delta_1(x) &= 0.1, \qquad \delta_2(x) = 0.5, \\ \delta_3(x) &= 0.1\,\mathbb{I}_{x=0} + 0.5\,\mathbb{I}_{x=1}, \quad \delta_4(x) = 0.5\,\mathbb{I}_{x=0} + 0.1\,\mathbb{I}_{x=1}. \end{aligned}$$

We assume in addition that the penalty for a wrong answer is 2 when $\theta = 0.1$ and 1 when $\theta = 0.5$. The *risk vectors* $(R(0.1, \delta), R(0.5, \delta))$ of the four estimators are then, respectively, $(0, 1)$, $(2, 0)$, $(0.2, 0.5)$, and $(1.8, 0.5)$. It is straightforward to see that the risk vector of any randomized estimator is a convex combination of these four vectors or, equivalently, that the *risk set*, $\mathcal{R}$, is the convex hull of the above four vectors, as represented by Figure 2.4.3.

In this case, the minimax estimator is obtained at the intersection of the diagonal of $\mathbb{R}^2$ with the lower boundary of $\mathcal{R}$. As shown by Figure 2.4.3, this estimator δ^* is randomized and takes the value $\delta_3(x)$ with probability $\alpha = 0.87$ and $\delta_2(x)$ with probability $1-\alpha$. The weight α is actually derived from the equation

$$0.2\alpha + 2(1 - \alpha) = 0.5\alpha.$$

This estimator δ^* is also a (randomized) *Bayes estimator* with respect to the prior

$$\pi(\theta) = 0.22\,\mathbb{I}_{0.1}(\theta) + 0.78\,\mathbb{I}_{0.5}(\theta);$$

the prior probability $\pi_1 = 0.22$ corresponds to the slope between $(0.2, 0.5)$ and $(2, 0)$, i.e.,

$$\frac{\pi_1}{1 - \pi_1} = \frac{0.5}{2 - 0.2}.$$

Notice that every randomized estimator that is a combination of δ_2 and of δ_3 is a Bayes estimator for this distribution, but that δ^* only is also a minimax estimator. ‖

Similar to minimax estimators, a least favorable distribution does not necessarily exist since its existence depends on a separating hyperplane theorem that does not always apply (see Pierce (1973), Brown (1976), Berger (1985a), and Chapter 8). In addition, Strawderman (1973) shows that, in the special case when $x \sim \mathcal{N}_p(\theta, I_p)$, there is no minimax proper Bayes estimator if $p \leq 4$. From a more practical point of view, Lemma 2.4.10 provides sufficient conditions of minimaxity.

Lemma 2.4.13 *If δ_0 is a Bayes estimator with respect to π_0 and if $R(\theta, \delta_0) \leq r(\pi_0)$ for every θ in the support of π_0, δ_0 is minimax and π_0 is the least favorable distribution.*

[3] The computations in this example are quite simple. See Chapter 8 for details.

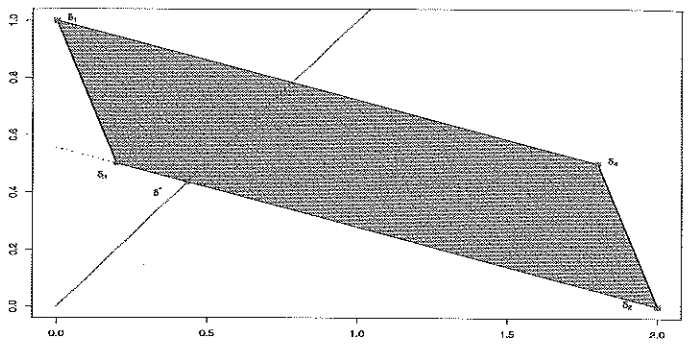

Figure 2.4.2. *Risk set for the estimation of the Bernoulli parameter.*

Example 2.4.14 (Berger (1985a)) Consider $x \sim \mathcal{B}(n, \theta)$ when θ is to be estimated under the quadratic loss,

$$\mathrm{L}(\theta, \delta) = (\delta - \theta)^2.$$

Bayes estimators are then given by posterior expectations (see Section 2.5) and, when $\theta \sim \mathcal{B}e\left(\frac{\sqrt{n}}{2}, \frac{\sqrt{n}}{2}\right)$, the posterior mean is

$$\delta^*(x) = \frac{x + \sqrt{n}/2}{n + \sqrt{n}}.$$

Moreover, this estimator has *constant risk*, $R(\theta, \delta^*) = 1/4(1+\sqrt{n})^2$. Therefore, integrating out θ, $r(\pi) = R(\theta, \delta^*)$ and δ^* is minimax according to Lemma 2.4.13. Notice the difference with the maximum likelihood estimator, $\delta_0(x) = x/n$, for the small values of n, and the unrealistic concentration of the prior around 0.5 for larger values of n. ‖

Since minimax estimators usually correspond to *generalized Bayes estimators*, it is often necessary to use a limiting argument to establish minimaxity, rather than computing directly the Bayes risk as in Lemma 2.4.13.

Lemma 2.4.15 *If there exists a sequence (π_n) of proper prior distributions such that the generalized Bayes estimator δ_0 satisfies*

$$R(\theta, \delta_0) \leq \lim_{n \to \infty} r(\pi_n) < +\infty$$

for every $\theta \in \Theta$, then δ_0 is minimax.

Example 2.4.16 When $x \sim \mathcal{N}(\theta, 1)$, the maximum likelihood estimator $\delta_0(x) = x$ is a generalized Bayes estimator associated with the Lebesgue measure on $\mathbb{R}$ and the quadratic loss. Since $R(\delta_0, \theta) = \mathbb{E}_\theta(x - \theta)^2 = 1$, this risk is the limit of the Bayes risks $r(\pi_n)$ when π_n is equal to $\mathcal{N}(0, n)$, as $r(\pi_n) = \frac{n}{n+1}$. Therefore, the maximum likelihood estimator δ_0 is minimax. Note that this argument can be extended directly to the case $x \sim \mathcal{N}_p(\theta, I_p)$ to establish that δ_0 is minimax for every p. ‖

When the space Θ is compact, minimax Bayes rules (or estimators) can be exactly described, owing to the *separated zeros principle* in complex calculus: if $R(\theta, \delta^\pi)$ is not constant and is analytic, the set of θ's where $R(\theta, \delta^\pi)$ is maximal is separated and, in the case of a compact set Θ, is necessarily finite.

Theorem 2.4.17 *Consider a statistical problem that simultaneously has a value, a least favorable distribution π_0, and a minimax estimator δ^{π_0}. Then, if $\Theta \subset \mathbb{R}$ is compact and if $R(\theta, \delta^{\pi_0})$ is an analytic function of θ, then either π_0 has a finite support or $R(\theta, \delta^{\pi_0})$ is constant.*

Example 2.4.18 Consider $x \sim \mathcal{N}(\theta, 1)$, with $|\theta| \leq m$, namely, $\theta \in [-m, m]$. Then, according to Theorem 2.4.17, least favorable distributions have necessarily a finite support, $\{\pm\theta_i,\ 1 \leq i \leq \omega\}$, with cardinal 2ω and supporting points θ_i depending on m. In fact, the only estimator with constant risk is $\delta_0(x) = x$, which is not minimax in this case. In general, the exact determination of n and of the points θ_i can only be done numerically. For instance, when $m \leq 1.06$, the prior distribution with weights $1/2$ at $\pm m$ is the *unique* least favorable distribution. Then, for $1.06 \leq m \leq 2$, the support of π contains $-m$, 0, and m. See Casella and Strawderman (1981) and Bickel (1981) for details, and Johnstone and MacGibbon (1992) for a similar treatment of the Poisson model. ‖

The above examples show why, while being closely related to the Bayesian paradigm, the minimax principle is not necessarily appealing from a Bayesian point of view. Indeed, apart from the fact that minimax estimators are sometimes randomized, as in Example 2.4.12, Examples 2.4.14 and 2.4.18 show that the least favorable prior is often unrealistic because it induces a strong prior bias towards a few points of the sample space. For Example 2.4.18, Gatsonis et al. (1987) have shown that uniform priors are good substitutes to the point mass priors, although they are not minimax.

Extensions of Theorem 2.4.17 to the noncompact case are given in Kempthorne (1988). In multidimensional settings, when the problem is invariant under rotation, the least favorable distributions are uniform on a sequence of embedded spheres (see Robert et al. (1990)). The practical problem of determining the points of the support is considered in Kempthorne (1987) and Eichenauer and Lehn (1989).

In settings where the problem has a value, it is often difficult to derive the least favorable distribution and alternative methods are then necessary to produce a minimax estimator. Chapter 9 shows how the exhibition of some invariance structures of the model may lead to identify the best equivariant estimator and a minimax estimator (Hunt–Stein Theorem). Unfortunately, the conditions under which this theorem applies are difficult to check and often do not hold.

Lastly, when a minimax estimator has been derived, its optimality is still to be assessed: there may exist several minimax estimators and some may

perform uniformly better than others. It is then necessary to introduce a second (and more local) criterion to compare minimax estimators, i.e., estimators that perform well globally.

2.4.4 Admissibility

This second frequentist criterion induces a partial ordering on $\mathcal{D}^*$ by comparing the frequentist risks of the estimators, $R(\theta, \delta)$.

Definition 2.4.19 *An estimator δ_0 is inadmissible if there exists an estimator δ_1 which dominates δ_0, that is, such that, for every θ,*

$$R(\theta, \delta_0) \geq R(\theta, \delta_1)$$

and, for at least one value θ_0 of the parameter,

$$R(\theta_0, \delta_0) > R(\theta_0, \delta_1).$$

Otherwise, δ_0 is said to be admissible.

This criterion is particularly interesting for its *reductive* action. Indeed, at least in theory, it seems logical to advocate that inadmissible estimators should not be considered at all since they can be uniformly improved. For instance, the Rao–Blackwell Theorem then implies that, for convex losses, randomized estimators are inadmissible. However, admissibility alone is not enough to validate the use of an estimator. For instance, constant estimators $\delta(x) = \theta_0$ are usually admissible because they produce the exact value at $\theta = \theta_0$. From a frequentist point of view, it is then important to look for estimators satisfying both optimalities, that is, minimaxity and admissibility. In this regard, two results can be mentioned.

Proposition 2.4.20 *If there exists a unique minimax estimator, this estimator is admissible.*

Proof. If δ^* is the only minimax estimator, for any estimator $\tilde{\delta} \neq \delta^*$,

$$\sup_\theta R(\theta, \tilde{\delta}) > \sup_\theta R(\theta, \delta^*).$$

Therefore, $\tilde{\delta}$ cannot dominate δ^*. □□

Notice that the converse to this result is false, since there can exist several minimax admissible estimators. For instance, in the $\mathcal{N}_p(\theta, I_p)$ case, there exist proper Bayes minimax estimators when $p \geq 5$ (Strawderman (1973) and Fourdrinier and Strawderman (1999)). When the loss function L is strictly convex (in d), it also allows for the following characterization.

Proposition 2.4.21 *If δ_0 is admissible with constant risk, δ_0 is the unique minimax estimator.*

Proof. For any $\theta_0 \in \Theta$, $\sup_\theta R(\theta, \delta_0) = R(\theta_0, \delta_0)$. Therefore, if there exists δ_1 such that $\bar{R} \leq \sup_\theta R(\theta, \delta_1) < R(\theta_0, \delta_0)$, δ_0 cannot be admissible. Similarly, if $\bar{R} = \sup_\theta R(\theta, \delta_1) = R(\theta_0, \delta_0)$ and if θ_1 is such that $R(\theta_1, \delta_1) < \bar{R}$,

δ_1 dominates δ_0. Therefore, when δ_0 is admissible, the only possible case is that there exists δ_1 such that $R(\theta, \delta_1) = R(\theta, \delta_0)$ for every $\theta \in \Theta$. And this is also impossible when δ_0 is admissible (see Exercise 2.36). □□

Again, notice that the converse of this result is false. There may be minimax estimators with constant risk that are inadmissible: actually, they are certainly inadmissible if there are other minimax estimators. For instance, this is the case for $\delta_0(x) = x$ when $x \sim \mathcal{N}_p(\theta, I_p)$ and $p \geq 3$ (see Note 2.8.2). There also are cases when there is no minimax admissible estimator (this requires that there is no *minimal complete class*, see Chapter 8).

The previous section showed that minimaxity can sometimes be considered from a Bayesian perspective as the choice by Nature of a maximin strategy (least favorable distribution), π, therefore that *some* minimax estimators are Bayes. Admissibility is even more strongly related to the Bayes paradigm in the sense that, in most statistical problems, the Bayes estimators are "spanning" the class of admissible estimators, i.e., the latter can be expressed as Bayes estimators or generalized Bayes estimators or limits of Bayes estimators. Chapter 8 deals in more detail with the relations between Bayes estimators and admissibility. We only give here two major results.

Proposition 2.4.22 *If a prior distribution π is strictly positive on Θ, with finite Bayes risk and the risk function, $R(\theta, \delta)$, is a continuous function of θ for every δ, the Bayes estimator δ^π is admissible.*

Proof. Suppose δ^π is inadmissible and consider δ' which uniformly dominates δ^π. Then, for every θ, $R(\theta, \delta') \leq R(\theta, \delta^\pi)$ and, in an *open set* C of Θ, $R(\theta, \delta') < R(\theta, \delta^\pi)$. Integrating out this inequality, we derive that

$$r(\pi, \delta') < r(\pi, \delta^\pi) = \int_\Theta R(\theta, \delta^\pi)\pi(\theta)\, d\theta,$$

which is impossible. □□

Proposition 2.4.23 *If the Bayes estimator associated with a prior π is unique, it is admissible.*

The proof of this result is similar to the proof of Proposition 2.4.20. Even if the Bayes estimator is not unique, it is still possible to exhibit at least *one* admissible Bayes estimator. When the loss function is strictly convex, the Bayes estimator is necessarily unique and thus admissible, according to the above proposition.

Example 2.4.24 (Example 2.4.14 continued) The estimator δ^* is a (proper) Bayes estimator, therefore admissible, and it has constant risk. Therefore, it is the *unique minimax estimator* under squared error loss. ‖

Notice that Proposition 2.4.22 contains the assumption that the Bayes risk is finite. Otherwise, every estimator is, in a way, a Bayes estimator (see

Exercise 2.43). On the other hand, some admissibility results can be established for improper priors. This is why we prefer to call *generalized Bayes* estimators the estimators associated with an infinite Bayes risk, rather those corresponding to an improper prior. This choice implies that the Bayes estimators of different quantities associated with the same prior distribution can be simultaneously regular Bayes estimators and generalized Bayes estimators, depending on what they estimate. This also guarantees that regular Bayes estimators will always be admissible, as shown by the following result.

Proposition 2.4.25 *If a Bayes estimator, δ^π, associated with a (proper or improper) prior π and a strictly convex loss function, is such that the Bayes risk,*

$$r(\pi) = \int_\Theta R(\theta, \delta^\pi)\pi(\theta)\, d\theta,$$

is finite, δ^π is admissible.

Example 2.4.26 Consider $x \sim \mathcal{N}(\theta, 1)$ and the null hypothesis $H_0 : \theta \leq 0$ is tested against the alternative hypothesis $H_1 : \theta > 0$. This testing problem is an *estimation* problem if we consider the estimation of the indicator function $\mathbb{I}_{H_0}(\theta)$. Under the quadratic loss

$$(\mathbb{I}_{H_0}(\theta) - \delta(x))^2,$$

we can propose the following estimator

$$\begin{aligned} p(x) &= P_0(X > x) \qquad (X \sim \mathcal{N}(0,1)) \\ &= 1 - \Phi(x), \end{aligned}$$

called the *p-value*, which is considered as a good frequentist answer to the testing problem (see Kiefer (1977) and Casella and Berger (1987)). Using Example 1.5.1, it is easy to show that p is a generalized Bayes estimator under Lebesgue measure and quadratic loss, since $\pi(\theta|x)$ is the $\mathcal{N}(x, 1)$ distribution and

$$\begin{aligned} p(x) &= \mathbb{E}^\pi[\mathbb{I}_{H_0}(\theta)|x] = P^\pi(\theta < 0|x) \\ &= P^\pi(\theta - x < -x|x) = 1 - \Phi(x). \end{aligned}$$

Moreover, the Bayes risk of p is finite (Exercise 2.34). Therefore, the p-value, when taken as an estimator of $\mathbb{I}_{H_0}$, is admissible. (See Section 5.4 for an extended analysis of the properties of the p-value.) ‖

Example 2.4.27 In the setting of the previous example, if θ is the parameter of interest, $\delta_0(x) = x$ is a generalized Bayes estimator under quadratic loss, but

$$\begin{aligned} r(\pi, \delta_0) &= \int_{-\infty}^{+\infty} R(\theta, \delta_0)\, d\theta \\ &= \int_{-\infty}^{+\infty} 1\, d\theta = +\infty. \end{aligned}$$

Therefore, Proposition 2.4.23 is useless in this case to assess the admissibility of δ_0. While δ_0 is actually admissible, its admissibility must be established through a sequence of proper priors, as shown in Chapter 8. ‖

Example 2.4.28 Consider $x \sim \mathcal{N}_p(\theta, I_p)$. If the parameter of interest is $||\theta||^2$ and the prior distribution is the Lebesgue measure on $\mathbb{R}^p$, since $\mathbb{E}^\pi[||\theta||^2|x] = \mathbb{E}[||y||^2]$, with $y \sim \mathcal{N}_p(x, I_p)$, the Bayes estimator under quadratic loss is

$$\delta^\pi(x) = ||x||^2 + p.$$

This generalized Bayes estimator is not admissible because it is dominated by $\delta_0(x) = ||x||^2 - p$ (Exercise 2.35). Since the classical risk is $R(\theta, \delta^\pi) = \text{var}(\|x\|^2) + 4p^2$, the Bayes risk is infinite. This phenomenon shows that the Lebesgue measure is not necessarily the best noninformative choice for a prior measure when the parameter of interest is a subvector of the parameter (see Chapter 3). ‖

2.5 Usual loss functions

When the setting of an experiment is such that the utility function cannot be determined (lack of time, limited information, etc.), a customary alternative is to resort to classical losses, which are mathematically tractable and well documented. Of course, this approach is an approximation of the underlying statistical model and should only be adopted when the utility function is missing. We conclude this section with a note on more intrinsic loss functions, although these are rarely used in practice. (See also Note 2.8.1 for a description of losses used in imaging.)

2.5.1 The quadratic loss

Proposed by Legendre (1805) and Gauss (1810), this loss is undoubtedly the most common evaluation criterion. Founding its validity on the ambiguity of the notion of *error* in statistical settings (i.e., measurement error versus random variation), it also gave rise to many criticisms, commonly dealing with the fact that the squared error loss

$$\mathrm{L}(\theta, d) = (\theta - d)^2 \tag{2.5.1}$$

penalizes large deviations too heavily.

However, convex loss functions like (2.5.1) have the incomparable advantage of avoiding the paradox of *risk lovers* and to exclude randomized estimators. Another usual justification for the quadratic loss is that it provides a Taylor expansion approximation to more complex symmetric losses (see Exercise 4.14 for a counterexample). In his 1810 paper, Gauss already acknowledged the arbitrariness of the quadratic loss and was defending it

on grounds of simplicity. Although the criticisms over a systematic use of the quadratic loss are quite valid, this loss is nonetheless extensively used because it gives intuitively sound Bayesian solutions, i.e., those one would naturally suggest as estimators for a non-decision-theoretic inference based on the posterior distribution. In fact, the Bayes estimators associated with the quadratic loss are the posterior means. However, note that the quadratic loss is not the only loss enjoying this property. Losses leading to posterior means as the Bayes estimators are called *proper losses* and characterized in Lindley (1985), Schervish (1989), van der Meulen (1992), and Hwang and Pemantle (1994). (See also Exercise 2.15.)

Proposition 2.5.1 *The Bayes estimator δ^π associated with the prior distribution π and with the quadratic loss* (2.5.1)*, is the posterior expectation*

$$\delta^\pi(x) = \mathbb{E}^\pi[\theta|x] = \frac{\int_\Theta \theta f(x|\theta)\pi(\theta)\,d\theta}{\int_\Theta f(x|\theta)\pi(\theta)\,d\theta}.$$

Proof. Since

$$\mathbb{E}^\pi[(\theta-\delta)^2|x] = \mathbb{E}^\pi[\theta^2|x] - 2\delta\mathbb{E}^\pi[\theta|x] + \delta^2,$$

the posterior loss actually attains its minimum at $\delta^\pi(x) = \mathbb{E}^\pi[\theta \mid x]$. □□

The following corollaries are straightforward to derive.

Corollary 2.5.2 *The Bayes estimator δ^π associated with π and with the weighted quadratic loss*

$$\mathrm{L}(\theta,\delta) = \omega(\theta)(\theta-\delta)^2, \tag{2.5.2}$$

where $\omega(\theta)$ is a nonnegative function, is

$$\delta^\pi(x) = \frac{\mathbb{E}^\pi[\omega(\theta)\theta|x]}{\mathbb{E}^\pi[\omega(\theta)|x]}.$$

Corollary 2.5.3 *When $\Theta \in \mathbb{R}^p$, the Bayes estimator δ^π associated with π and with the quadratic loss,*

$$\mathrm{L}(\theta,\delta) = (\theta-\delta)^t Q(\theta-\delta),$$

is the posterior mean, $\delta^\pi(x) = \mathbb{E}^\pi[\theta|x]$, for every positive-definite symmetric $p \times p$ matrix Q.

Corollary 2.5.2 exhibits a (weak) *duality* between loss and prior distribution, in the sense that it is equivalent to estimate θ under (2.5.2) with the prior π, or under (2.5.1) with the prior $\pi_\omega(\theta) \propto \pi(\theta)\omega(\theta)$. Moreover, while admissibility is independent of the weight factor, the Bayes estimator strongly depends on the function ω. For instance, δ^π may not exist if ω increases too fast to $+\infty$. On the other hand, Corollary 2.5.3 shows that the Bayes estimators are robust with respect to the quadratic form Q. (Shinozaki (1975) has also proved that admissibility does not depend on Q.)

The quadratic loss is particularly interesting in the setting of bounded parameter spaces when the choice of a more subjective loss is impossible. In fact, this loss is quite tractable and the approximation error is usually negligible. Indeterminacy about the loss function (and thus its replacement by a quadratic approximation) often occurs in *accuracy evaluation*, including for instance *loss estimation* (see Rukhin (1988a,b), Lu and Berger (1989a,b), Hwang, Casella et al. (1992), Robert and Casella (1993, 1994), and Fourdrinier and Wells (1994)).

Example 2.5.4 (Example 2.4.9 continued) We are looking for an evaluation of the performances of the estimator

$$\delta(x_1,x_2) = \begin{cases} \dfrac{x_1+x_2}{2} & \text{if } x_1 \neq x_2, \\ x_1+1 & \text{otherwise,} \end{cases}$$

by $\alpha(x_1,x_2)$ under the quadratic criterion

$$[\mathbb{I}_\theta(\delta(x_1,x_2)) - \alpha(x_1,x_2)]^2,$$

where $\mathbb{I}_\theta(v)$ is 1 if $v=\theta$, 0 otherwise; the function α somehow evaluates the probability that δ takes the true value θ. (This is a special case of loss estimation, when the loss function is $1-\mathbb{I}_\theta(\delta)$.) Two estimators can be proposed:

(i) $\alpha_0(x_1,x_2) = 0.75$, which is the expectation of $\mathbb{I}_\theta(\delta(x_1,x_2))$; and

(ii) $\alpha_1(x_1,x_2) = \begin{cases} 1 & \text{if } x_1 \neq x_2, \\ 0.50 & \text{if } x_1 = x_2. \end{cases}$

The risks of the two evaluators are then

$$\begin{aligned} R(\theta,\alpha_0) &= \mathbb{E}_\theta\left(\mathbb{I}_\theta(\delta(x_1,x_2)) - 0.75\right)^2 \\ &= 0.75 - (0.75)^2 = 0.1875\,; \end{aligned}$$

and

$$\begin{aligned} R(\theta,\alpha_1) &= \mathbb{E}_\theta\left(\mathbb{I}_\theta(\delta(x_1,x_2)) - \alpha_1(x_1,x_2)\right)^2 \\ &= (0.5)^2\frac{1}{2} = 0.125\,. \end{aligned}$$

Therefore, α_1 is a better estimator of the performances of δ than α_0. As mentioned in Berger and Wolpert (1988), this domination result is quite logical and it suggests that a conditional evaluation of estimators is more appropriate. ‖

2.5.2 The absolute error loss

An alternative solution to the quadratic loss in dimension one is to use the absolute error loss,

$$\mathrm{L}(\theta,d) = |\,\theta - d\,|, \tag{2.5.3}$$

already considered by Laplace (1773) or, more generally, a multilinear function

$$\text{(2.5.4)} \qquad \mathrm{L}_{k_1,k_2}(\theta,d) = \begin{cases} k_2(\theta-d) & \text{if } \theta > d\,, \\ k_1(d-\theta) & \text{otherwise.} \end{cases}$$

Such functions increase more slowly than the quadratic loss. Therefore, while remaining convex, they do not overpenalize large but unlikely errors. Huber (1964) also proposed a mixture of the absolute error loss and the quadratic loss, in order to keep a quadratic penalization around 0,

$$\tilde{\mathrm{L}}(\theta,d) = \begin{cases} (d-\theta)^2 & \text{if } \mid d-\theta \mid < k, \\ 2k \mid d-\theta \mid - k^2 & \text{otherwise.} \end{cases}$$

Although a convex[4] loss, the mixed loss slows down the progression of the quadratic loss for large errors and has a robustifying effect. Unfortunately, there usually is no explicit derivation of Bayes estimators under this loss $\tilde{\mathrm{L}}$.

Proposition 2.5.5 *A Bayes estimator associated with the prior distribution π and the multilinear loss* (2.5.4) *is a $(k_2/(k_1+k_2))$ fractile of $\pi(\theta|x)$.*

Proof. The following classical equality

$$\begin{aligned} \mathbb{E}^{\pi}[\mathrm{L}_{k_1,k_2}(\theta,d)|x] &= k_1\int_{-\infty}^{d}(d-\theta)\pi(\theta|x)\,d\theta + k_2\int_{d}^{+\infty}(\theta-d)\pi(\theta|x)\,d\theta \\ &= k_1\int_{-\infty}^{d}P^{\pi}(\theta<y|x)\,dy + k_2\int_{d}^{+\infty}P^{\pi}(\theta>y|x)\,dy, \end{aligned}$$

is obtained by an integration by parts. Taking the derivative in d, we get

$$k_1 P^{\pi}(\theta<d|x) - k_2 P^{\pi}(\theta>d|x) = 0,$$

i.e.,

$$P^{\pi}(\theta<d|x) = \frac{k_2}{k_1+k_2}.$$

□□

In particular, if $k_1 = k_2$, i.e., in the case of the absolute error loss, the Bayes estimator is the posterior median, which is the estimator obtained by Laplace (see Example 1.2.4). Notice that, when π has a nonconnected support, Proposition 2.5.5 provides examples of multiple Bayes estimators for some values of x (see Exercise 2.40).

2.5.3 The $0-1$ loss

This loss is mainly used in the classical approach to hypothesis testing, as formalized by Neyman and Pearson (see Section 5.3). More generally, this

[4] Again, if we insist so much on *convexity*, it is because it ensures that randomized estimators are suboptimal from a frequentist point of view. Therefore, a statistical decision-theoretic approach that would agree as much as possible with the Likelihood Principle necessarily calls for convex losses. This requirement obviously eliminates bounded losses.

is a typical example of a nonquantitative loss. In fact, for this loss, the penalty associated with an estimate δ is 0 if the answer is correct and 1 otherwise.

Example 2.5.6 Consider the test of $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \notin \Theta_0$. Then $\mathcal{D} = \{0,1\}$, where 1 stands for acceptance of H_0 and 0 for rejection (in other words, the function of θ to be estimated is $\mathbb{I}_{\Theta_0}(\theta)$). For the $0-1$ loss, i.e.,

$$\text{(2.5.5)} \qquad \mathrm{L}(\theta,d) = \begin{cases} 1-d & \text{if } \theta \in \Theta_0 \\ d & \text{otherwise,} \end{cases}$$

the associated risk is

$$\begin{aligned} R(\theta,\delta) &= \mathbb{E}_{\theta}[\mathrm{L}(\theta,\delta(x))] \\ &= \begin{cases} P_{\theta}(\delta(x)=0) & \text{if } \theta \in \Theta_0\,, \\ P_{\theta}(\delta(x)=1) & \text{otherwise,} \end{cases} \end{aligned}$$

which are exactly the *type–one and type–two errors* underlying the Neyman–Pearson theory. ‖

This loss is not very interesting because of its nonquantitative aspect, and we will consider in Chapter 5 some alternative theories for testing hypotheses. The associated Bayes estimators also reflect the primitive aspect of such a loss (see also Exercise 2.41).

Proposition 2.5.7 *The Bayes estimator associated with π and with the loss* (2.5.5) *is*

$$\delta^{\pi}(x) = \begin{cases} 1 & \text{if } P(\theta \in \Theta_0|x) > P(\theta \notin \Theta_0|x), \\ 0 & \text{otherwise,} \end{cases}$$

i.e., $\delta^{\pi}(x)$ is equal to 1 if and only if $P(\theta \in \Theta_0|x) > 1/2$.

2.5.4 Intrinsic losses

It may occur that some settings are so noninformative that not only the loss function is unknown, but there is not even a natural parameterization. Such cases happen when the distribution $f(x|\theta)$ itself is of interest, for instance, in prediction settings.

However, as we mentioned in the previous section, the choice of the parameterization is important because, contrary to the maximum likelihood estimation approach, if g is a one-to-one transformation of θ, the Bayes estimator of $g(\theta)$ is usually different from the transformation by g of the Bayes estimator of θ under the same loss (see Exercise 2.36). This lack of invariance, although often troubling to beginners, is not usually a concern for decision-makers because it shows how the Bayesian paradigm can adapt to the estimation problem at hand *and* the selected loss function, while maximum likelihood estimation is totally loss-blind. But the few cases where

loss function and natural parameterization are completely unavailable may call for this kind of ultimate invariance. (See Wallace and Boulton (1975) for another approach.)

In such noninformative settings, it seems natural to use losses that compare directly the distributions $f(\cdot|\theta)$ and $f(\cdot|\delta)$ associated with the true parameter θ and the estimate δ. Such loss functions,

$$\mathrm{L}(\theta,\delta) = d(f(\cdot|\theta), f(\cdot|\delta)),$$

are indeed parameterization-free. Two usual distribution distances are

(1) the *entropy distance*

$$\mathrm{L_e}(\theta,\delta) = \mathbb{E}_\theta\left[\log\left(\frac{f(x|\theta)}{f(x|\delta)}\right)\right], \tag{2.5.6}$$

which is also called the Kullback–Leibler divergence and which is not a distance in the mathematical sense because of its asymmetry; and

(2) the *Hellinger distance*

$$\mathrm{L_H}(\theta,\delta) = \frac{1}{2}\mathbb{E}_\theta\left[\left(\sqrt{\frac{f(x|\delta)}{f(x|\theta)}} - 1\right)^2\right]. \tag{2.5.7}$$

Example 2.5.8 Consider $x \sim \mathcal{N}(\theta, 1)$. Then we have

$$\begin{aligned}\mathrm{L_e}(\theta,\delta) &= \frac{1}{2}\mathbb{E}_\theta[-(x-\theta)^2 + (x-\delta)^2] = \frac{1}{2}(\delta-\theta)^2,\\ \mathrm{L_H}(\theta,\delta) &= 1 - \exp\{-(\delta-\theta)^2/8\}.\end{aligned}$$

Considering the normal case when $\pi(\theta|x)$ is a $\mathcal{N}(\mu(x),\sigma^2)$ distribution, it is straightforward to show that the Bayes estimator is $\delta^\pi(x) = \mu(x)$ in both cases. ‖

The Hellinger loss is undoubtedly more intrinsic than the entropy loss, if only because it always exists (note that (2.5.7) is bounded above by 1). Unfortunately, while leading to explicit expressions of $\mathrm{L_H}(\theta,\delta)$ for the usual distribution families, it does not allow for an explicit derivation of the Bayes estimators, except in the special case treated above. On the contrary, in *exponential families*, the entropy loss provides explicit estimators which are the posterior expectations for the estimation of the *natural parameter* (see Chapter 3). Moreover, although quite different from the Hellinger loss, the entropy loss provides similar answers for the usual distribution families (see Robert (1996b)). There are also various theoretical reasons to defend the use of the Kullback-Leibler distance, ranging from information theory (Exercise 2.48) to the relevance of logarithmic scoring rule and the location-scale invariance of the distance, as detailed in Bernardo and Smith (1994).

2.6 Criticisms and alternatives

Some criticisms about the frequentist notions of minimaxity and admissibility have been mentioned in the previous sections. These concepts are actually of secondary interest from a purely Bayesian point of view, since, on one hand, admissibility is automatically satisfied by most Bayes estimators. On the other hand, minimaxity is somehow incompatible with the Bayesian paradigm, since, under a prior distribution, each value of the parameter cannot be equally weighted. However, minimaxity may be relevant from a robustness point of view, that is, when the prior information is not precise enough to determine the prior distribution.

It may happen that the decision-maker cannot define a loss function exactly. For instance, when the decision-maker is a committee comprising several experts, it is often the case that they differ about the relevant loss function (and sometimes even about the prior distribution). Starting with Arrow (1951), the literature on these extensions of classical Decision Theory is quite extensive (see Genest and Zidek (1986), Rubin (1987), and Van Eeden and Zidek (1993) for details and references).

When the loss function has not been completely determined, it might be assumed to belong to a parametrized class of loss functions, the decision maker selecting the most accurate parameter. Apart from L_p losses, two other possible classes are

$$\mathrm{L}_1(\theta,\delta) = \log(\alpha||\theta-\delta||^2+1), \qquad \mathrm{L}_2(\theta,\delta) = 1-\exp\{-c||\theta-\delta||^2\}.$$

An alternative approach more in tune with the Bayesian paradigm is to consider that, since the loss is partly unknown, this uncertainty can be represented by using a *random loss* $\mathrm{L}(\theta,\delta)$. The evaluation of estimators is then done by integrating out with respect to this additional variable: If F is the distribution of the loss, the objective function to minimize (in δ) is

$$\int_\Theta \int_\Omega \mathrm{L}(\theta,\delta,\omega)dF(\omega)\,d\pi(\theta|x), \tag{2.6.1}$$

where F possibly depends on θ or even on x. This case is actually the only interesting extension because, otherwise, to minimize (2.6.1) is equivalent to using the average loss

$$\bar{\mathrm{L}}(\theta,\delta) = \int_\Omega \mathrm{L}(\theta,\delta,\omega)\,dF(\omega).$$

Another approach to the lack of precision on the loss function consists of considering simultaneously a set of losses and look for estimators performing well for all these losses. Obviously, this multidimensional criterion only induces a *partial* ordering on estimators.

Example 2.6.1 Consider $x \sim \mathcal{N}_p(\theta, I_p)$. The parameter θ is estimated under quadratic loss. If the loss matrix Q is not exactly determined, a robust alternative is to include the losses associated with the matrices Q such that $Q_1 \preceq Q \preceq Q_1$ (where $A \preceq B$ means that the matrix $B - A$ is

nonnegative definite). Notice that, according to Corollary 2.5.3, the Bayes estimator is the same for all Q's. ‖

Example 2.6.2 In the setting of the above example, Brown (1975) shows that a shrinkage estimator of the form $(1-h(x))x$ dominates $\delta_0(x) = x$ for a class of quadratic losses, i.e., a class of matrices Q if and only if

$$\text{tr}(Q) - 2\lambda_{\max}(Q) > 0 \tag{2.6.2}$$

for every matrix in the class (where $\lambda_{\max}$ denotes the largest eigenvalue). Notice that this condition excludes the case $p \leq 2$, where δ_0 is actually admissible. The constant $\text{tr}(Q) - 2\lambda_{\max}(Q)$ also appears in the majorization constant of $||x||^2 h(||x||^2)$ (see Theorem 2.8.1). Therefore, (2.6.2) is both a necessary and sufficient condition for the Stein effect to occur. ‖

The ultimate criterion in loss robustness is called *universal domination* and was introduced in Hwang (1985). It actually takes into account the set of all losses $\ell(||\delta - \theta||_Q)$, for a given norm $||x||_Q = x^t Q x$ and all nondecreasing functions ℓ. An estimator δ_1 will be said to *universally dominate* another estimator δ_2 if, for every ℓ,

$$\mathbb{E}_\theta[\ell(||\delta_1(x) - \theta||_Q)] \leq \mathbb{E}_\theta[\ell(||\delta_2(x) - \theta||_Q)].$$

A second criterion is called *stochastic domination*: δ_1 *stochastically dominates* δ_2 if, for every $c > 0$,

$$P_\theta(||\delta_1(x) - \theta||_Q \leq c) \geq P_\theta(||\delta_2(x) - \theta||_Q \leq c).$$

Although this criterion seems more intrinsic and less related to Decision Theory than universal domination, Hwang (1985) has shown that the two criteria are actually equivalent.

Theorem 2.6.3 *An estimator δ_1 universally dominates an estimator δ_2 if and only if δ_1 stochastically dominates δ_2 .*

Proof. The estimator δ_1 stochastically dominates δ_2 if, for every $c > 0$,

$$P_\theta(||\delta_1(x) - \theta||_Q \leq c) \geq P_\theta(||\delta_2(x) - \theta||_Q \leq c).$$

This can be rewritten as

$$\mathbb{E}_\theta\left[\mathbb{I}_{[c,+\infty[}(||\delta_1(x) - \theta||_Q)\right] \leq \mathbb{E}_\theta\left[\mathbb{I}_{[c,+\infty[}(||\delta_2(x) - \theta||_Q)\right].$$

Since $\ell(t) = \mathbb{I}_{[c,+\infty[}(t)$ is a nondecreasing function of t, universal domination implies stochastic domination. The converse follows from the fact that the first moments of two stochastically ordered random variables are also ordered. □□

Moreover, these two criteria are not empty since Hwang (1985) has established the following domination result: If $x \sim \mathcal{T}_\alpha(\mu, \sigma^2)$, Student's t-distribution with α degrees of freedom, some shrinkage estimators universally dominate $\delta_0(x) = x$. If the dimension is not too small (usually, $p = 4$ is sufficient), Brown and Hwang (1989) virtually showed that, if $x \sim \mathcal{N}_p(\theta, \Sigma)$, the estimator $\delta_0(x)$ is admissible for universal domination *if and only if* $Q = \Sigma$. For other choices of the matrix Q and p large enough, δ_0 is stochastically dominated. Therefore, even though this criterion is less discriminating than usual losses, it allows for comparison, and even for a Stein effect, since classical estimators are not necessarily optimal.

The study of multiple losses is not very developed from a Bayesian point of view, since Bayes estimators usually vary with a change in the loss function. However, in a very special case, Rukhin (1978) has shown that the Bayes estimators were *independent of the loss function*. Under some regularity assumptions, this case corresponds to the equation

$$\log f(x|\theta) + \log \pi(\theta) = A_1(x)e^{\alpha\theta} + A_2(x)e^{-\alpha\theta} + A_3(x),$$

where π is the prior distribution. Therefore, for this *exponential family* (see Section 3.3.3),

$$f(x|\theta) = \frac{B(x)}{\pi(\theta)} \exp\{A_1(x)e^{\alpha\theta} + A_2(x)e^{-\alpha\theta}\}, \tag{2.6.3}$$

the Bayes estimators are *universal*, because they do not depend on the loss. The next chapter covers in detail the case of exponential families, which are classes of distributions on $\mathbb{R}^k$ with densities

$$f(x|\theta) = c(\theta)h(x)\exp[R(\theta)\cdot T(x)],$$

where $R(\theta), T(x) \in \mathbb{R}^p$. However, notice that (2.6.3) is a rather special exponential family.

2.7 Exercises

Section 2.2

2.1 Show that, if the utility function U is convex, every $P \in \mathcal{P}_\mathcal{E}$ satisfies

$$\mathbb{E}^P[r] = \int_\mathcal{R} r\, dP(r) \preceq P.$$

Conclude that a concave loss is not realistic.

2.2 Consider four dice with respective numbers on their faces $(4,4,4,4,0,0)$, $(3,3,3,3,3,3)$, $(6,6,2,2,2,2)$, $(1,1,1,5,5,5)$. Two players roll one die each and compare their outcome. Show that the relation die $[i]$ beats die $[j]$ is intransitive, i.e., that for every choice of the first player the second player can choose a die so that the probability of winning is greater than 0.5. Relate this example to the Pitman closeness setting of Note 2.8.3.

2.3 Show that $\mathcal{P}_\mathcal{B} \subset \mathcal{P}_\mathcal{E}$, i.e., that bounded reward distributions have a finite expected utility.

2.4 Show Lemmas 2.2.2 and 2.2.3.

2.5 *(DeGroot (1970)) In order to show the extension of Theorem 2.2.4 from $\mathcal{P}_\mathcal{B}$ to $\mathcal{P}_\mathcal{E}$, consider a sequence s_m decreasing (for $\preceq$) in $\mathcal{R}$ such that, for every