Distributional inference

Albers, Casper Johannes

Publication date:
2003

Citation for published version (APA):
Albers, C. J. (2003). *Distributional inference: the limits of reason*. s.n.

Part I

The limits of reason

Chapter 1

How to assign probabilities if you must

> *'Without the information the model was clear: a random drawing from 1,...,6. After the information one is completely muddled. Giving out information should be accompanied by the knowledge of the information-policy.'*
>
> J. Hemelrijk[1]

Empirical evidence can sometimes be incorporated in a probabilistic analysis by conditioning with respect to the observations. Usually, the underlying probability distribution and also the conditional distribution are not completely known. The assignment of probabilities will then require a compromise. The making of such compromise goes beyond mathematical theory: a statistical discussion is needed. It depends on the context whether the result of such discussion is almost compelling, reasonable, or not really agreeable. This is illustrated by means of a simple example from the area of predictive distributional inference.

This chapter is an extended version of the article with the same title in Statistica Neerlandica (Albers and Schaafsma, 2001b) (2001b), and the technical report with the same title (Albers, 2000) (2000).

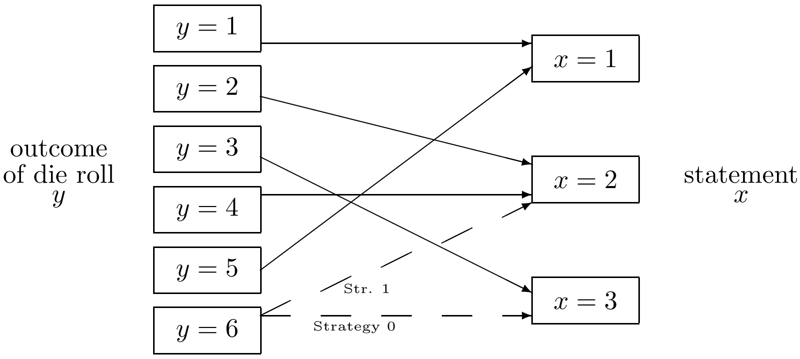
1.1 Predictive distributional inference, an example

Most theories of probability, Bayesian statistics included, prescribe to incorporate empirical evidence by computing conditional distributions. The availability of such prescriptions suggests that the approach is compelling. This may be the case if a fair die is rolled once and somebody has been instructed to tell us, without lying, whether the number of eyes y is even or odd, but the compellingness disappears if the instructions are less specific. If somebody provides us with the information that y is even, it can be very misleading to infer that the probabilities of the possibilities $y = 2,\ 4$ and 6 are equal to $\frac{1}{3}$.

The example

A fair die has been rolled once and the true number of eyes y has been made available to some person (or Nature), henceforth referred to as Player I. Player I has to provide Player II (the statistician) with true information x about y. He has to choose one of the statements made in Figure 1.1. Note that there is a difficulty if $y = 6$. In that case, Player I has to choose between $x = 2$ and $x = 3$.

[1] Rules for building statistical models, *Statistica Neerlandica*, **32**:3, 1978.



$x = 1$: y is neither even, nor a triple
$x = 2$: y is even
$x = 3$: y is a triple

Figure 1.1: Visualization of the outcome space

Question
Suppose Player I provides us with the information x (either 1, 2, or 3). Which probabilities $q_x(\eta)$ should be assigned to the theoretical possibilities η for y?

Preliminary exploration
If $x = 1$, then the theoretical possibilities are 1 and 5 and nobody will criticize the specification $q_1(1) = q_1(5) = \frac{1}{2}$ because the die is fair. If $x = 2$ we are in trouble because if we assume that Player I chooses $x = 2$ in case $y = 6$ (Strategy 1) then $q_2(2) = q_2(4) = q_2(6) = \frac{1}{3}$ is logical, but if we assume that he would choose $x = 3$ (Strategy 0) then $x = 2$ implies that $y = 6$ is impossible and that $q_2(2) = q_2(4) = \frac{1}{2}$ is appropriate . If $x = 3$, a similar discussion can be made. Hemelrijk (1978) states that 'Information about the result of an experiment is only reliable if the receiver knows the information-policy used, i.e. if he knows which information would be given for every possible result of the experiment'. The nonexistence of a unique solution, due to an unknown information-policy, in the cases $x = 2$ and $x = 3$ implies that answering the question is only possible in the sense of making a compromise.

1.2 Background information

The situation now obtained is characteristic for almost everything in statistics: the solution depends on an unknown true value, here strategy number $i \in \{0, 1\}$. We shall discuss a logical approach, a Bayesian approach and we give a preview of the Fisher-Neyman-Pearson-Wald approach we recommend. This section will be concluded with some additional comments.

The situation in the preliminary exploration above is such that it is logical to specify that, e.g., $q_2(6) \in \{0, \frac{1}{3}\}$ or, more precisely, that $q_2(6)$ is equal to 0 if $i = 0$ and equal to $\frac{1}{3}$ if $i = 1$. The latter statement is logically valid, but useless because we do not know whether $i = 0$ or $i = 1$. Simply stating that $q_2(6)$ is either 0 or $\frac{1}{3}$ is a possibility which is in line with some theories in logic. The interesting book by Cleave (1991) starts with Carnap's statement that 'in logic, the science of valid inference, there are no morals'. Indeed, if one refuses to choose between 0 and $\frac{1}{3}$ then no moral principles will be needed. The statistician, however, will accept the task of specifying a number $q_2(6)$, at least if a 'reasonable' compromise is possible.

The Bayesians adhere to the perspective that Player I chooses Strategy 1 with probability ρ. As a consequence $q_2(6)$ will be chosen equal to $\mathrm{P}(Y = 6|X = 2) = \rho/(2+\rho)$. The question now is *'which ρ?'*. In this problem, the choice $\rho = \frac{1}{2}$ cannot be defended on the basis of symmetry arguments. The Bayesian might continue by simply using $\rho = \frac{1}{2}$ because this 'is in the middle' of the interval $[0, 1]$ of possibilities, or he might continue by simply using ρ itself as the outcome of a uniformly distributed random variable. Such constructions are not necessarily reasonable. That is why we abandon the Bayesian approach and start from scratch in Section 1.3 with a 'classical' statistical approach where i is regarded as an unknown number. In Section 1.5 we accept the idea that i is the outcome of a random variable such that the possibility 1 has probability ρ. The discussion leads to the conclusion that this does not help much if one knows nothing about the true value of ρ.

The latter situation is not very realistic in the sense that if one knows that i is the outcome of a random variable then one will usually also have some information about ρ. Ignoring such prior information, computations to be discussed in the sequel lead to $\rho = .495$ (and, hence $q_2(6) = .198$) as the (Bayesian) solution, which corresponds to the minimax regret procedure. The minimax risk procedure is characterized by $\rho = .667$ (and $q_2(6) = .250$). A referee of Albers and Schaafsma (2001b) remarked that, without moral information the choice of principle cannot be discussed. We agree to a certain extent: many statisticians have experienced that Wald's minimax risk principle is often too conservative and that it may even lead to degenerate results such that, e.g., the observations are completely ignored (see, e.g., Schaafsma, 1969). Such experiences suggest that the choice of ρ 'should be' closer to .495 than to .667.

The reason we pay much attention to this elementary and impractical example is that we are interested in the foundations of predictive inference, predictive distributional inference in particular. Our example is a very simple example of a problem where a predictive inference is required. That our problem is, in some sense, one step too simple, will be shown and dealt with in Section 1.7.

Note that the whole perspective of the example may change drastically if additional information is provided. If, e.g., Player II observes that Player I flips a coin before issuing the statement $x = 2$ then he should *not* conclude that $\rho = \frac{1}{2}$ and, hence $q_2(6) = \frac{1}{2}/(2 + \frac{1}{2}) = .20$ is appropriate. It is 'logical' to conclude that Player I has seen a six, and hence $q_2(6) = 1$, because in any other case it makes no sense to flip a coin.

With these preliminarities in mind, the reader will, hopefully, appreciate the following discussions illustrating

1. Fisher's desire to create an inductive logic;
2. Popper's statement that *induction is a myth*;
3. The fact from life that *induction is a must.*

Epistẽmẽ (infallible knowledge about the universe) is beyond reach but we can do our best in providing 'approximations'. We are aware of the fact that, in some sense, we have done our best too much. The reader can restrict himself to Sections 1.3, 1.4, 1.5, and 1.9.

1.3 A Fisher-Neyman-Pearson-Wald approach

Fisher proposed methods of inference while Neyman and Pearson, and especially Wald, tried to make comparative analyses of such methods in the hope that some method comes out best (from certain perspectives). The first and very essential step in this approach is to shift the attention from the concrete situation with a given statement x (either 1,2, or 3) to that where probabilities $q_\xi(\eta)$ have to be assigned for each value ξ a priori possible for x. It seems reasonable to restrict the attention to the class

$$\mathcal{D} = \left\{ Q_{a,b} \mid \tfrac{1}{3} \le a \le \tfrac{1}{2},\ \ \tfrac{1}{2} \le b \le 1 \right\}$$

of procedures $Q_{a,b}$ defined in Table 1.1. One of the arguments behind this restriction is that in a Bayesian context (Strategy 1 is chosen with probability ρ) the posterior probabilities $\mathrm{P}(Y = \eta | X = \xi)$ are of this kind, with $a = (2+\rho)^{-1}$ and $b = (2-\rho)^{-1}$. Note that in this section we do *not* make the assumption that Strategy i is the outcome of a random variable.

Simply choosing $q_2(2) = q_2(4) = q_2(6) = \frac{1}{3}$ and $q_3(3) = q_3(6) = \frac{1}{2}$ is a possibility, but not a clever one. It corresponds to $Q_{1/3,1/2}$ and is represented by the left-lower point A of the rectangle in the left graph of Figure 1.2. If Player I is known to act according to Strategy 1 then $Q_1 = Q_{1/3,1}$ is the procedure to choose. It is represented as the left-upper point of the rectangle. If he would choose Strategy 0 then $Q_0 = Q_{1/2,1/2}$ (the right-lower point) is 'logically valid' from the probabilistic viewpoint. As we are unaware of Player I's strategy and yet are forced to assign probabilities, a *compromise* will be needed. Averaging the parameters of Q_0 and Q_1

Table 1.1: Procedures $Q_{a,b}$

$q_\xi(\eta)$	$\eta = 1$	$\eta = 2$	$\eta = 3$	$\eta = 4$	$\eta = 5$	$\eta = 6$
$\xi = 1$	$\frac{1}{2}$	0	0	0	$\frac{1}{2}$	0
$\xi = 2$	0	a	0	a	0	$1-2a$
$\xi = 3$	0	0	b	0	0	$1-b$

we obtain $Q_{5/12,3/4}$ (point E). This solution is not satisfactory from an intellectual viewpoint: it is like cutting a Gordian knot without examining it. To improve this situation, we adopt the perspective of the theory of statistical decision functions. In its usual form this theory tries to prescribe how statements should be made about true values of unknown parameters (here strategy number i). Our situation is different in the sense that a predictive statement is required, namely about the true value of some random variable Y. The distributional form of inference is more natural if such predictive inference is required than in the classical situation. On the other hand, it is more complicated to arrive at satisfactory results (unless a Bayesian approach is adopted). We, for example, do not know whether the fundamental Wald-Lehmann minimal-complete class theorem is valid in predictive inference at large. In our special example it will follow from the concrete analysis that the minimal complete class corresponds to the class of all Bayes procedures. We, however, doubt whether this holds in general, an extension in Section 1.7 did not provide an answer. In this respect it is interesting to quote Eaton (1999, p. 851), who claims

> 'The more stringent evaluation of predictive distributions using decision theoretic notions (minimaxity, admissibility, etcetera) has received very little attention in the literature. A few results can be found in Eaton (1982, 1992) but a body of work providing hard evidence - that is, i.e. theorems - that specific predictive distributions will perform well in particular situations is, in the main, lacking.'

Anyway, given the observation x, we have to choose a probability distribution $Q = Q(x)$ on $\{1, \ldots, 6\}$ with probabilities $q_x(1), \ldots, q_x(6)$ and think in terms of the loss $L(y, Q)$ to be incurred if the true value y is revealed. An important requirement is that the loss is *proper*: if y is the outcome of a random variable Y with its probabilities $p(\eta) = \mathrm{P}(Y = \eta)$ known, then L is said to be proper if

$$\mathbf{E}\, L(Y, Q) = \sum_{\eta=1}^{6} L(\eta, Q) p(\eta)$$

is minimum, as a function of Q, if the corresponding probabilities satisfy $q(\eta) = p(\eta)$. At this moment the elaborations are restricted to the logarithmic loss function

$$L(y, P(x)) = -\log(q_x(y)).$$

The properness of this loss function is an immediate consequence of the positiveness of the Kullback-Leibler information number because, if P denotes the true distribution of Y with $\mathrm{P}(Y = \eta) = p(\eta)$, and Q is any other distribution with $Q(\{\eta\}) = q(\eta)$, then

$$\mathbf{E}\, L(Y, Q) - \mathbf{E}\, L(Y, P) = \sum_{\eta=1}^{6} \log\left(\frac{p(\eta)}{q(\eta)}\right) p(\eta)$$

is positive if $P \neq Q$ (Good, 1952). Using this loss function we can determine the risk (expected loss) of the procedure $Q_{a,b}$. As the distribution of (X, Y) depends on

$\mathrm{P}_i(X=\xi, Y=\eta)$	$\eta=1$	$\eta=2$	$\eta=3$	$\eta=4$	$\eta=5$	$\eta=6$
$\xi=1$	1/6	0	0	0	1/6	0
$\xi=2$	0	1/6	0	1/6	0	$i/6$
$\xi=3$	0	0	1/6	0	0	$(1-i)/6$

Table 1.2: Distributions P_i

whether Player I chooses Strategy 0 (saying $x = 3$ if $y = 6$) or Strategy 1 (saying $x = 2$), there are two distributions involved. They are given in Table 1.2.

We shall have to consider the risks (expected losses)

$$\begin{aligned} R(0, Q_{a,b}) &= -\tfrac{1}{6}\log\left(\left(\tfrac{1}{2}\right)^2 a^2 b(1-b)\right) \\ R(1, Q_{a,b}) &= -\tfrac{1}{6}\log\left(\left(\tfrac{1}{2}\right)^2 a^2 b(1-2a)\right), \end{aligned}$$

obtained from Tables 1.1 and 1.2. $R(0, Q_{a,b})$ is minimum if both a^2 and $b(1-b)$ $((a,b) \in [\frac{1}{3},\frac{1}{2}] \times [\frac{1}{2},1])$ are maximum, i.e. if $Q_0 = Q_{1/2,1/2}$ is used. Similarly $R(1, Q_{a,b})$ is minimum if $Q_1 = Q_{1/3,1}$ is used. The fact that the minimization of these risks as a function of a can be separated from the minimization as a function of b indicates that our example is not representative for predictive inference at large. The minimum risk achieved by using the best procedure is called the envelope risk. It depends on i and is given by

$$\begin{aligned} R^*(0) &= R(0, Q_0) = \log 2 \\ R^*(1) &= R(1, Q_1) = \tfrac{1}{3}\log 2 + \tfrac{1}{2}\log 3. \end{aligned}$$

The risk of any procedure $Q_{a,b}$ is larger than the envelope risk. The difference is referred to as the regret or shortcoming. Using $S(i, Q_{a,b})$ as notation, we have

$$\begin{aligned} S(0, Q_{a,b}) &= R(0, Q_{a,b}) - R^*(0) = -\tfrac{1}{6}\log\left(16a^2 b(1-b)\right) \\ S(1, Q_{a,b}) &= R(1, Q_{a,b}) - R^*(1) = -\tfrac{1}{6}\log\left(27a^2 b(1-2a)\right) \end{aligned}$$

Note that $S(0, Q_{a,b})$ is a decreasing function of a and an increasing function of b if $(a,b) \in [\frac{1}{3},\frac{1}{2}] \times [\frac{1}{2},1]$. For $S(1, Q_{a,b})$ the situation is reversed. In our example, the minimal complete class of procedures corresponds to the class of Bayes procedures. This class can be obtained by minimizing the convex combination

$$\begin{aligned} &(1-\rho)S(0, Q_{a,b}) + \rho S(1, Q_{a,b}) = -\tfrac{1}{6}\left(2\log a + \rho\log(1-2a) + \right. \\ &\left. \log b + (1-\rho)\log(1-b) + (1-\rho)\log 2^4 + \rho\log 3^3\right) \end{aligned}$$

of both shortcomings. With elementary calculus it can be seen that this linear combination is minimum as a function of a and b if $a = (2+\rho)^{-1}$ and $b = (2-\rho)^{-1}$.

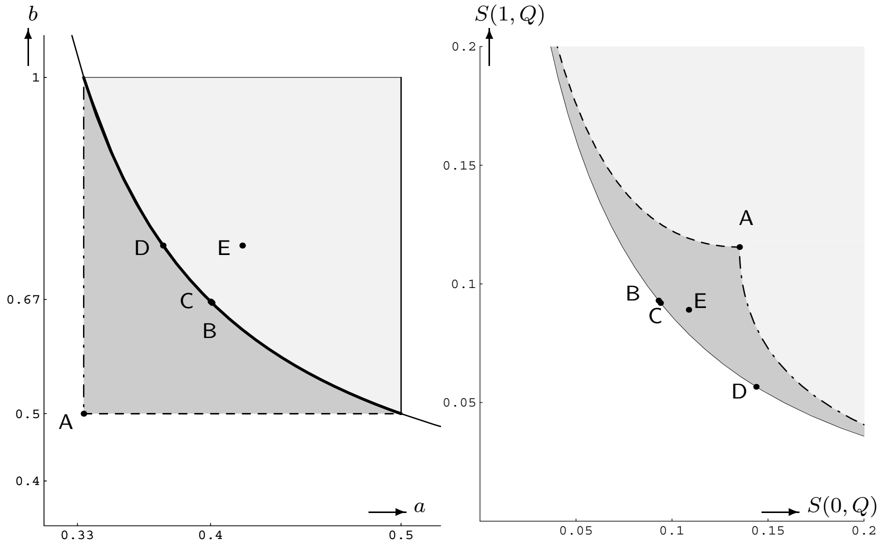


Figure 1.2: Visualization of the procedures in the (a, b)-plane (on the left) and the $(S(0, Q), S(1, Q))$-plane (right). The solid curves correspond to the procedures Q_ρ. Only the part inside the box in the figure to the left deserves consideration. The dashed lines correspond to procedures $Q_{a,1/2}$, the dot-dashed ones to $Q_{1/3,b}$. Explanation of the points A, B, C, D and E will be given in the text, especially at the end of Section 1.3.

This is expressed by $b = a/(4a - 1)$ and by the boldfaced curve in Figure 1.2 (left). Henceforth we use the notation

$$Q_\rho = Q_{\frac{1}{2+\rho}, \frac{1}{2-\rho}}$$

to denote the procedure of the form $Q_{a,b}$ that minimizes $(1-\rho)S(0, Q_{a,b}) + \rho S(1, Q_{a,b})$. Note that $Q_0 = Q_{1/2,1/2}$ and $Q_1 = Q_{1/3,1}$ are as before. Looking for a compromise $Q_{a,b}$ between Q_0 and Q_1, one should not go (too far) beyond the boldfaced curve in Figure 1.2 (left) which characterizes the procedures which are 'admissible' in the sense of Wald's theory. For the general case we have

$$\begin{aligned} S(0, Q_\rho) &= -\tfrac{1}{6} \log \left(16 \frac{1-\rho}{(2+\rho)^2 (2-\rho)^2} \right) \\ S(1, Q_\rho) &= -\tfrac{1}{6} \log \left(27 \frac{\rho}{(2+\rho)^3 (2-\rho)} \right). \end{aligned}$$

The corresponding points constitute the left-lower bound of the regret set

$$\mathcal{S} = \left\{ (S(0, Q_{a,b}), S(1, Q_{a,b})) \mid \tfrac{1}{3} \le a \le \tfrac{1}{2},\ \tfrac{1}{2} \le b \le 1 \right\}$$

which is, of course, the risk set when the origin is shifted to $(R^*(0), R^*(1))$. In this example the minimal complete class corresponds to the class of all admissible procedures as well as to the class of all Bayes procedures.

1.4 A detailed discussion of some specific procedures

Figure 1.2 provides visualizations of the parameters (a, b) of the procedures $Q_{a,b}$ (left) and of the corresponding points $S((0, Q_{a,b}), S(1, Q_{a,b}))$ (right). Table 1.3 gives details about the following specific points.

Point A

The naive procedure $Q_{1/3,1/2}$ has already been discussed at the beginning of the previous section. It corresponds to the idea that the probabilities $q_2(2) = q_2(4) = q_2(6) = \frac{1}{3}$ have to be assigned if $x = 2$ (because 2, 4 and 6 are equiprobable if Strategy 1 is chosen) and that the probabilities $q_3(3) = q_3(6) = \frac{1}{2}$ have to be assigned if $x = 3$ (because 3 and 6 are possible and equiprobable if Strategy 0 is used). The snake in the grass is that if $x = 2$, and Strategy 0 would have been chosen in case $y = 6$, then 2, 4 and 6 are not equiprobable at all, because $y = 6$ is impossible. A similar argument holds for $x = 3$, if Strategy 1 would have been chosen in case $y = 6$. It is very difficult for some probabilists to accept that the information *'the die is fair, the number of eyes is even'* does not imply that the outcomes 2, 4 and 6 are equiprobable. In the present context such a probabilist will be less unwilling to deviate from that what he regards as the foundations of Probability Theory than in, e.g., the quiz-master's problem. The reason is that we can present a precise analysis of the situation. The crux is, of course, that *the source of the information ('the number of eyes is even') has to be made part of the probabilistic model.* In the present Fisher-Neyman-Pearson-Wald approach this is done by referring to the true but unknown number i of the strategy which Player I would apply if he would have been confronted by $y = 6$. It follows from the discussion that the naive procedure $Q_{1/3,1/2}$ represented by A is not appropriate: a should be larger than $\frac{1}{3}$ and b larger than $\frac{1}{2}$. How much larger, that is the question.

Table 1.3: Specialization of some regret points

Procedure		Description	$S(0,Q)$	$S(1,Q)$
A	$Q_{1/3,1/2}$	naive conditional probabilities	.135	.116
B	$Q_{.495}$	minimax regret: $S(0,Q_\rho) = S(1,Q_\rho)$	.093	.093
C	$Q_{1/2}$	Bayes with uniform prior	.094	.092
D	$Q_{2/3}$	minimax risk: $R(0,Q_\rho) = R(1,Q_\rho)$	.144	.057
E	$Q_{5/12,3/4}$	naive compromise	.109	.089
	$Q_{2/5}$		.072	.119
	$Q_{.55}$		.107	.080

Point B
We are attracted by the idea to minimize the maximum regret (see Section 1.2). This can be achieved by looking for the Bayes procedure Q_ρ which satisfies $S(0, Q_\rho) = S(1, Q_\rho)$. The computations provide $\rho = .495$, etcetera and thus lead to procedure $Q_{.4007,.6648}$. See Table 1.3 and Figure 1.2 (point B) and notice that both regrets .093 are smaller than those of the naive procedure.

Point C
The Bayesian approach has a considerable appeal but it requires the choice of the prior probability ρ which, in the context of this section, is a fictitious construct (see Section 1.5 for a different possibility). We can ignore probabilistic terminology by simple stating that we want to minimize some weighted average of the risks, or, equivalently, of the regrets, $\frac{1}{2}S(0, Q_{a,b}) + \frac{1}{2}S(1, Q_{a,b})$. This is then achieved by the procedure $Q_{1/2}$ represented by the point C. Note that one coordinate of the corresponding regret point is larger, and one is smaller than that of B.

Point D
WALD (1947) was fascinated by the theory of games as presented in VON NEUMANN AND MORGENSTERN (1944) . This leads to minimizing the maximum risk procedure, which is obtained by equating the two risks. This provides $b = 2a$ and corresponds to point D.

Point E
At the beginning of this section we argued that a compromise will be needed and we suggested the algebraically natural candidate $Q_{5/12,3/4}$ is not satisfactory from an intellectual viewpoint. Indeed, $(\frac{5}{12}, \frac{3}{4})$ is beyond the boldfaced curve in Figure 1.2 (left) which corresponds to the admissible procedures. This generates the task to construct a procedure with both regrets decreased. Our first try was to solve $1/(2 + \rho) = \frac{5}{12}$ which provides $\rho = \frac{2}{5}$, the regret $S(1, Q_{2/5})$ being larger than that of the naive compromise. Solving $1/(2 - \rho) = \frac{3}{4}$ leads to the minimax risk procedure. Our final try was to take $\rho = .55$ (thus $Q_{.3922,.6897}$), which indeed has smaller regrets than the naive compromise.

A vexed issue at the end
The parameter ρ was used as a technical device to generate the minimal complete class $\{Q_\rho \mid 0 \leq \rho \leq 1\}$ of Bayes procedures. The procedure $Q_{1/3,1/2}$ was much too naive in the sense that it assumes that Strategy 1 has been chosen if $x = 2$ and Strategy 0 if $x = 3$. This does not necessarily correspond to the facts. It cannot be denied, however, that the outcome x contains some information with respect to Player I's choice of strategy. If we try to exploit this information by choosing ρ depending on x, then the procedure obtained may be reasonable but it will not be admissible in the sense of Wald, the admissibility requires that the procedure is in the class $\{Q_\rho \mid 0 \leq \rho \leq 1\}$.

1.5 What if Player I uses a randomized strategy?

As complete consensus can obviously not be achieved, additional knowledge would be welcome. In this section we restrict the attention to the idea that Player I chooses Strategy 1 with probability ρ, this 'physical' probability being either fully known (Situation 1) or fully unknown (Situation 2). In the end of this section, we will discuss the third possibility of partial knowledge.

Situation 1 may appear if we are involved in a two-person zero-sum game. Knowledge of the pay-off matrix will then not only affect the choice of ρ (if one accepts the minimax principle) but also the restriction to the class $\mathcal{D} = \{Q_{a,b} \mid \frac{1}{3} \leq a \leq \frac{1}{2}, \frac{1}{2} \leq b \leq 1\}$. This makes it clear that the 'solutions' presented in Section 1.4 are only reasonable if further information is absent.

Henceforth the attention is concentrated on Situation 2: the randomization probability ρ will then appear as the unknown true value of the parameter $\theta \in \Theta = [0,1]$, and the factual pair (x, y) is the outcome of a pair (X, Y) of random variables with distribution uniquely determined by ρ. As nothing is known about ρ, it is intuitively clear that the additional information is not worth much. We introduce random variables (X_θ, Y_θ) having the distribution P_θ which (X, Y) would have had, given $\rho = \theta$. If i is replaced by θ in Table 1.2 then one obtains a table of $\mathrm{P}(X_\theta = \xi, Y_\theta = \eta)$ values.

The risk $-\mathbf{E}\ \log(q_X(Y))$ depends on the true value ρ of θ. In general we have

$$\begin{aligned} R(\theta, Q_{a,b}) &= -\mathbf{E}\ \log(q_{X_\theta}(Y_\theta)) \\ &= -\tfrac{1}{6}(\log(2^{-2}a^2 b) - \theta \log(1-2a) - (1-\theta)\log(1-b)). \end{aligned}$$

Note that this risk is equal to $(1-\theta)R(0, Q_{a,b}) + \theta R(1, Q_{a,b})$ and, hence, is given by $R(0, Q_{a,b})$ and $R(1, Q_{a,b})$ in the end points.

For fixed θ, the risk $R(\theta, Q_{a,b})$ is minimum if $(a, b) = (1/(2+\theta), 1/(2-\theta))$. The procedure Q_θ thus obtained is Bayes with respect to all prior distributions τ on $[0,1]$ which have θ as their expectation. The envelope risk is

$$\begin{aligned} R^*(\theta) &= R(\theta, Q_\theta) \\ &= (1-\theta)R(0, Q_\theta) + \theta R(1, Q_\theta) \end{aligned}$$

where $R(0, Q_\theta)$ is obtained from $R(0, Q_{a,b})$ by substituting

$$(a, b) = \left(\frac{1}{2+\theta}, \frac{1}{2-\theta}\right)$$

and $R(1, Q_\theta)$ follows from $R(1, Q_{a,b})$. The envelope risk displayed in Figure 1.3, is such that $R^*(0)$ and $R^*(1)$ correspond to $R^*(0) = \log 2$ and $R^*(1) = \frac{1}{3}\log 2 + \frac{1}{2}\log 3$. The maximum is reached for $\theta = \frac{2}{3}$; the minimax risk procedure is $Q_{2/3}$. The shortcoming $S(\theta, Q_{a,b}) = R(\theta, Q_{a,b}) - R^*(\theta)$ is a convex function of θ because $R^*(\theta)$ is concave and $R(\theta, Q_{a,b})$ is linear in θ. As a consequence the minimax regret and the

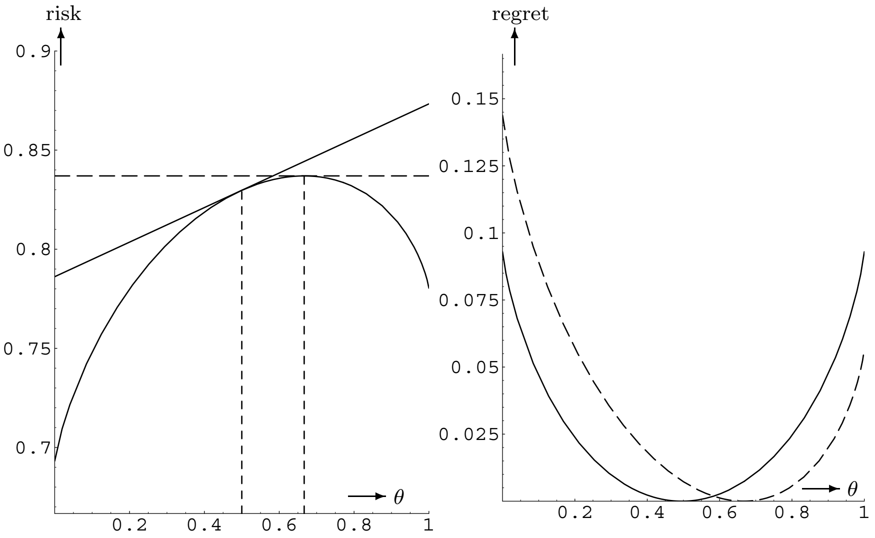


Figure 1.3: The randomized strategy game. On the left: envelope risk $R^*(\theta)$ (curve), minimax risk procedure (dashed line) and minimax regret procedure (solid line). On the right, the same procedures, but now with the regret along the vertical axis.

minimax procedure are exactly the same as in Section 1.4. The area under the regret function is minimized by using $Q_{1/2} = Q_{2/5,2/3}$ which, thus, is the minimal average regret procedure. The main difference with Figure 1.2 is that shortcoming functions are now visualized as functions of θ.

The elaboration about randomized strategies suggests that nothing is gained if we know that Player I chooses Strategy 1 with some completely unknown probability ρ. One might argue that this state of ignorance changes as soon as the outcome x is available. The underlying random variable X assumes the values $1, 2$, or 3 with probabilities $\frac{1}{3}$, $\frac{1}{3} + \frac{1}{6}\rho$, $\frac{1}{3} - \frac{1}{6}\rho$ respectively. Shouldn't this information be used to replace the a priori choice of $\rho = \frac{1}{2}$ and $a = \frac{2}{5}$, $b = \frac{2}{3}$ by an a posteriori choice of $\rho = \frac{1}{2}$, $\frac{1}{2} + \epsilon$, $\frac{1}{2} - \epsilon$ if $x = 1, 2, 3$? In practice, Bayesian statisticians adapt their prior if it is in conflict with actual data. In the present example it is easy to see that this approach leads to a procedure that is 'inadmissible' from a theoretical viewpoint: the x-dependent choice of ρ suggested corresponds to a $Q_{a,b}$ with

$$(a, b) = \left(\frac{1}{\frac{5}{2} + \epsilon}, \frac{1}{\frac{3}{2} + \epsilon} \right)$$

not in the arc of admissible procedures.

In practice, the additional information will not be of the form of the two extremes

$q_\xi(\eta)$	$\eta = 1$	$\eta = 2$	$\eta = 3$	$\eta = 4$	$\eta = 5$	$\eta = 6$
$\xi = 1$	c	0	0	0	$(1-c)$	0
$\xi = 2$	0	a	0	d	0	$(1-a-d)$
$\xi = 3$	0	0	b	0	0	$(1-b)$

Table 1.4: Procedures $Q_{a,b,c,d}$

('ρ fully known' or 'ρ completely unknown') suggested in the beginning of Section 1.5. A confidence interval or distributional inference about ρ may be available. This may affect the conclusion that $Q_{a,b}$ with $(a, b) \approx (0.40, 0.67)$ is the rule to choose. That a compromise solution may be affected if additional information is provided, is completely natural though it illustrates the hazards involved in any discussion of the types presented.

1.6 Adapting the theory to alternative loss functions

Our theory started from the class $\mathcal{D}$ of procedures and concentrated the attention on risks and regrets based on logarithmic loss. It resulted in the opinion that (a, b) should be close to the arc

$$\left\{(a,b) \mid (a - \tfrac{1}{4})(b - \tfrac{1}{4}) = \tfrac{1}{16}\right\}$$

because these points generate the class $\{Q_\rho \mid 0 \le \rho \le 1\}$ of all Bayes procedures which corresponds to the minimal complete class. An extensive discussion suggested that the procedure to be chosen should be of the form Q_ρ where ρ is not too much different from $\frac{1}{2}$. The agreement among the probabilities actually assigned is such that none of these probabilities is completely compelling in the cases $x = 2$ or $x = 3$. They, however, are not unreasonable because the agreement between, e.g., $q_2(6) = .198$ (minimax average risk or regret) and $q_2(6) = .250$ (minimax risk) is quite satisfactory.

The theory was based on the restriction to the class $\mathcal{D}$ of procedures $Q_{a,b}$ (see the beginning of Section 1.2). Instead of $\mathcal{D}$ one might consider the more general class of procedures of the form $Q_{a,b,c,d}$ defined in Table 1.4. If one uses a *proper* loss function (see the end of this section) this extension is useless because the class $\mathcal{D}$ is essentially complete, in fact $\{Q_\rho : 0 \le \rho \le 1\} \subset \mathcal{D}$ is already (minimal) complete.

The reason for this is as follows. To derive the Bayes procedure w.r.t. some τ on $\Theta = [0, 1]$, let us invoke a probabilistic model where (x, y) appears as the outcome of a pair (X_τ, Y_τ) of random variables. The distribution of this pair corresponds to the marginal distribution of $(\tilde{T}, \tilde{X}, \tilde{Y})$ where $\mathcal{L}\,(\tilde{T}) = \tau$ and $\mathcal{L}\,((\tilde{X}, \tilde{Y}) \mid \tilde{T} = \theta) = P_\theta =$

$\mathcal{L}\,(X_\theta, Y_\theta)$. The Bayes risk is

$$\begin{aligned}
r(\tau, Q_{a,b,c,d}) &= \int R(\theta, Q_{a,b,c,d})\ \mathrm{d}\tau(\theta) \\
&= \int \mathbf{E}\ L(Y_\theta, Q_{a,b,c,d}(X_\theta))\ \mathrm{d}\tau(\theta) \\
&= \mathbf{E}\ L(\tilde{Y}, Q_{a,b,c,d}(\tilde{X})) \\
&= \mathbf{E}\ \Big(\mathbf{E}\ L(\tilde{Y}, Q_{a,b,c,d}(\tilde{X}))|\tilde{X}\Big)
\end{aligned}$$

The properness of L implies that we have to use the procedure with parameters a, b, c, and d, such that

$$Q_{a,b,c,d}(x) = \mathcal{L}\,(\tilde{Y}|\tilde{X} = x)$$

But the marginal distribution of $(\tilde{X}, \tilde{Y})$ has already been described. It is that of $(X_\tau, Y_\tau) = (X_\theta, Y_\theta)$ where $\theta = \int_0^1 u\ \mathrm{d}\tau(u)$. Hence

$$\mathcal{L}\,(\tilde{Y}|\tilde{X} = x) = Q_{\frac{1}{2+\theta},\frac{1}{2-\theta}}(x).$$

Another 'basis' of our theory was the usage of the logarithmic loss function. The question arises whether the agreement will be affected if the loss function is replaced by another (proper) one. Let us consider three proper loss functions: logarithmic, Brier (1950) and Epstein (1969) loss, respectively

$$\begin{aligned}
L_{\log}(y, Q(x)) &= -\log q_x(y), \\
L_{\mathrm{B}}(y, Q(x)) &= (1 - q_x(y))^2 + \sum_{\eta \neq y} (q_x(\eta))^2, \\
L_{\mathrm{E}}(y, Q(x)) &= \sum_{\eta=1}^{6} \left(\mathbf{1}_{\{y,\ldots,6\}}(\eta) - \sum_{\nu=1}^{\eta} q_x(\nu)\right)^2.
\end{aligned}$$

For the seven possible combinations of x and y, the losses incorporated are as in Table 1.5 From these losses, we can compute the risk functions $R_{\cdot}(0, Q_{a,b})$ and $R_{\cdot}(1, Q_{a,b})$ for Strategies 0 and 1 respectively, as well as the risk functions $R_{\cdot}(\theta, Q_{a,b})$ corresponding to the randomized strategy.

$$\begin{aligned}
R_{\log}(\theta, Q_{a,b}) &= -\tfrac{1}{6}\left(\log(2^{-2}a^2 b) - \theta \log(1 - 2a) - (1 - \theta)\log(1 - b)\right) \\
R_{\mathrm{B}}(\theta, Q_{a,b}) &= \tfrac{1}{6}\left(6(2 + \theta)a^2 - 12a + 2(2 - \theta)b^2 - 4b + 7\right) \\
R_{\mathrm{E}}(\theta, Q_{a,b}) &= \tfrac{1}{6}\left(10(2 + \theta)a^2 - 20a + 3(2 - \theta)b^2 - 6b + 11\right)
\end{aligned}$$

Due to the linear relation $\mathrm{P}_\theta = (1 - \theta)\mathrm{P}_0 + \theta \mathrm{P}_1$ and properness of the loss functions, the different envelope risks are obtained when the same procedures are used, as was

(x,y)	logarithmic	Brier	Epstein
$(1,1)$	$\log(2)$	0.5	1
$(1,5)$	$\log(2)$	0.5	1
$(2,2)$	$-\log(a)$	$6a^2-6a+2$	$10a^2-12a+4$
$(2,4)$	$-\log(a)$	$6a^2-6a+2$	$10a^2-8a+2$
$(2,6)$	$-\log(1-2a)$	$6a^2$	$10a^2$
$(3,3)$	$-\log(b)$	$2(1-b)^2$	$3(1-b)^2$
$(3,6)$	$-\log(1-b)$	$2b^2$	$3b^2$

Table 1.5: 7×3-matrix of possible losses

already noted before. These procedures are of the form $Q_\theta = Q_{(2+\theta)^{-1},(2-\theta)^{-1}}$ with endpoints $Q_{\frac{1}{2},\frac{1}{2}}$ and $Q_{\frac{1}{3},1}$ if $\theta = 0, 1$. The envelope risk functions are

$$\begin{aligned}
R^*_{\log}(\theta,Q_\theta) &= -\tfrac{1}{6}(-\log 2^2-((2+\theta)\log(2+\theta)+(2-\theta)\log(2-\theta)) \\
&\quad +(\theta\log\theta+(1-\theta)\log(1-\theta))) \\
R^*_{\mathrm{B}}(\theta,Q_\theta) &= \tfrac{1}{6}\left(7+\tfrac{1}{2+\theta}+\tfrac{5}{2-\theta}-\tfrac{28}{(2+\theta)(2-\theta)}\right) \\
R^*_{\mathrm{E}}(\theta,Q_\theta) &= \tfrac{1}{6}\left(11+\tfrac{1}{2+\theta}+\tfrac{8}{2-\theta}-\tfrac{44}{(2+\theta)(2-\theta)}\right)
\end{aligned}$$

The shortcomings, or regrets, are obtained by subtracting the envelope risks from the risks. Thus, $S(\theta,Q_\rho) = R(\theta,Q_\rho) - R(\theta,Q_\theta)$ for $\theta,\rho \in [0,1]$. The regret functions for our three different loss functions are

$$\begin{aligned}
S_{\log}(\theta,Q_\rho) &= -\tfrac{1}{6}(-\theta\log\tfrac{\rho}{(2+\theta)^3(2-\theta)}+(1-\theta)\log\tfrac{1-\rho}{(2+\theta)^2(2-\theta)^2}+\theta\log\theta+ \\
&\quad +(1-\theta)\log(1-\theta)-(2-\theta)\log(2-\theta)-(2+\theta)\log(2+\theta)) \\
S_{\mathrm{B}}(\theta,Q_\rho) &= \tfrac{(\rho-\theta)^2/6}{(2+\rho)^2(2-\rho)^2(2+\theta)(2-\theta)}((8-2\theta)\rho^2+(-16+2\theta)\rho+(16-4\theta)) \\
S_{\mathrm{E}}(\theta,Q_\rho) &= \tfrac{(\rho-\theta)^2/6}{(2-\rho)^2(2+\rho)^2(2-\theta)(2+\theta)}((26-7\theta)\rho^2+ \\
&\quad (-56+52\theta)\rho+(104-28\theta))
\end{aligned}$$

These formulas are sufficient material for a graphical analysis of the procedures following from the three different loss functions. We concentrate the attention on the motivation behind the five points A, ..., E considered before. In Table 1.6, for each of these points the corresponding procedure Q and regret points $S(0,Q), S(1,Q)$ are given.

A graphical representation can be seen in Figure 1.4 where, from above to below, two graphs for logarithmic, Brier and Epstein loss are displayed. The five points, along with the curve $\{Q_\rho|\rho\in[0,1]\}$ are plotted. The three left graphs are plotted in the (a,b)-plane, where the rectangle bounds the procedures that deserve consideration.

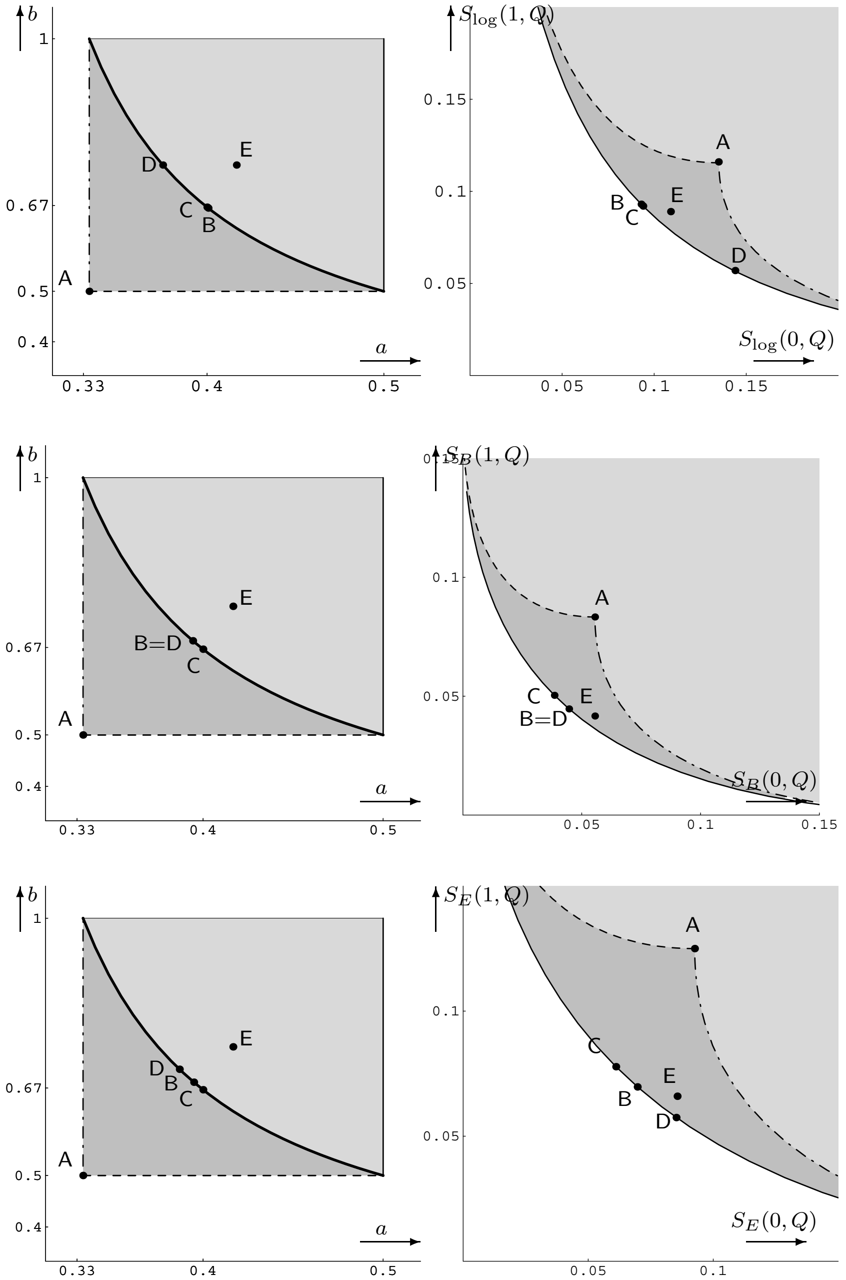


Figure 1.4: Visualization of logarithmic, Brier and Epstein scoring rules.

Table 1.6: Overview of regret points

	Logarithmic			Brier			Epstein		
	Q	$\theta = 0$	$\theta = 1$	Q	$\theta = 0$	$\theta = 1$	Q	$\theta = 0$	$\theta = 1$
A	$Q_{\frac{1}{3},\frac{1}{2}}$	.135	.116	$Q_{\frac{1}{3},\frac{1}{2}}$	.0556	.0833	$Q_{\frac{1}{3},\frac{1}{2}}$	.0926	.1250
B	$Q_{.496}$	.093	.093	$Q_{.536}$	.0447	.0447	$Q_{.532}$	.0697	.0697
C	$Q_{\frac{1}{2}}$	.094	.092	$Q_{\frac{1}{2}}$	.0385	.0504	$Q_{\frac{1}{2}}$	.0611	.0778
D	$Q_{\frac{2}{3}}$	.144	.057	$Q_{.536}$	.0447	.0447	$Q_{.584}$	.0852	.0575
E	$Q_{\frac{5}{12},\frac{3}{4}}$	.109	.089	$Q_{\frac{5}{12},\frac{3}{4}}$	.0556	.0417	$Q_{\frac{5}{12},\frac{3}{4}}$	.0856	.0660

The three right graphs present the $(S(0, Q_{a,b}), S(1, Q_{a,b}))$-points, part of the rectangle is mapped and folded.

In Figure 1.5 we have made a display of the risk for the case of logarithmic loss. Figure 1.6 provides a similar display, this time of the regret. For Brier and Epstein loss similar visualizations can be made. These figures display a similar behaviour. The left part of Figure 1.5 shows us the envelope risk $R^*_{\log}(\theta)$ (curve), the minimax risk procedure $R_{\log}(\theta, Q_{2/3})$ (dashed line) and the minimax regret procedure $R_{\log}(\theta, Q_{.4957})$ (solid line). In left of Figure 1.6 we see the same procedures. In the right part of Figure 1.5 the riskplane $R_{\log}(\theta, Q_\rho)$ $([\theta, \rho] \in (0,1)^2)$ is displayed. Notice that for fixed ρ, the risk is linear in θ. The three solid lines correspond to the envelope risk $(\theta = \rho)$ and minimax risk $(\rho = \frac{2}{3})$ and minimax regret $(\rho = 0.4957)$ procedures. The right part of Figure 1.6 shows a similar display, now in the regretplane. The figures on the left of Figures 1.5 and 1.6 can thus be described as a two-dimensional display of the right figures ('ignoring' the value of ρ).

Conclusions

The procedures corresponding to A, C and E are independent of the choice of loss function (as long as this loss function is proper), as was to be expected. Of course, the corresponding risks and regrets do shift. Note that in the case of Brier-loss, the values of the minimax risk and minimax regret procedures coincide because, in this case, the envelope risks in 0 and 1 are equal. The position of the minimax risk and minimax regret procedures are different, but remain in the form Q_θ. The differences are also small, all procedures suggest a strategy Q_θ with θ close to or in $[\frac{1}{2}, \frac{2}{3}]$. If the loss function is proper, then the class $\{Q_\rho | \rho : 0 \leq \rho \leq 1\}$, of Bayes rules (they are also Somewhere Minimal Risk), will not be affected, $Q_{1/2}$ will minimize the average risk or regret (if a randomized strategy is considered, integration should be with respect to Lebesgue measure). The position of the ρ values corresponding to the minimax regret or the minimax risk procedures will become somewhat different.

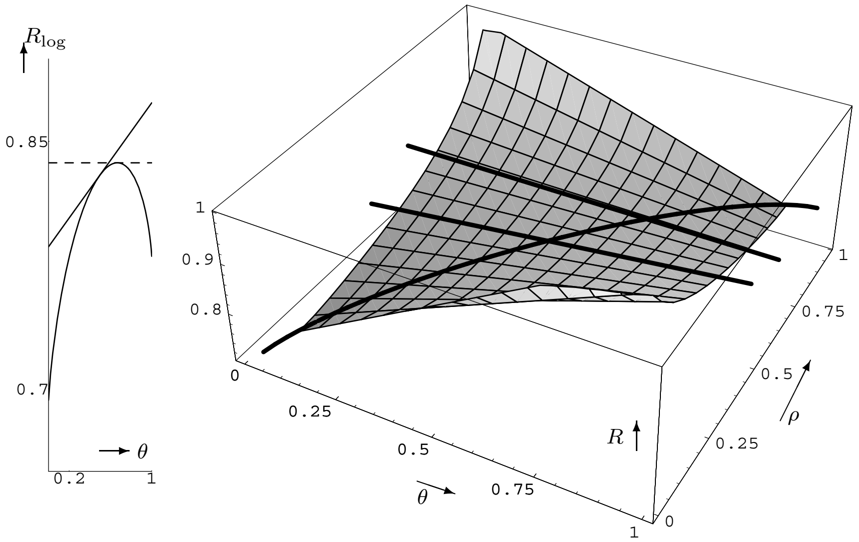


Figure 1.5: Risks for logarithmic loss.

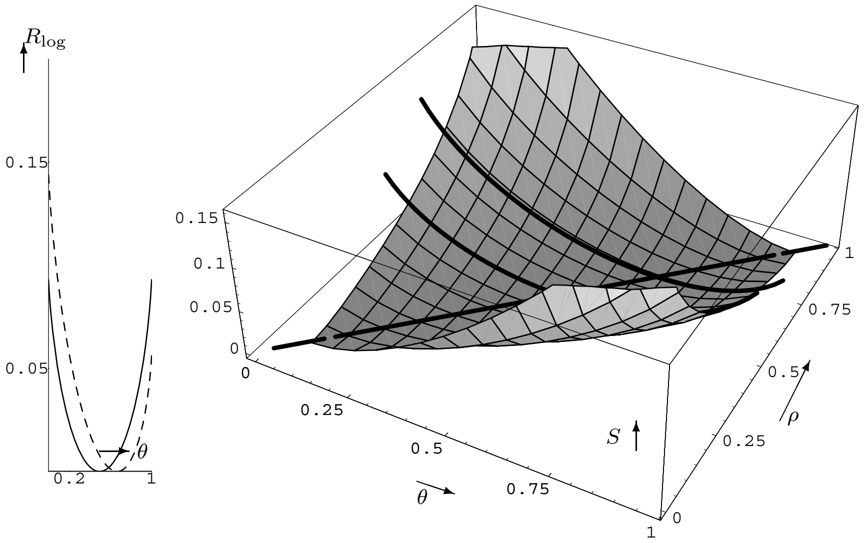


Figure 1.6: Regrets for logarithmic loss.

1.7 Extension

This and the next section are concerned with generalizations to and extensions of the example discussed so far. The reader can continue directly with Section 1.9 without loss of understanding of the main ideas of this chapter. The example discussed so far is in some sense paradigmic for statistics at large. In another sense it is not sufficiently general to be called paradigmic because the linearity

$$\mathrm{P}_\theta = (1-\theta)\mathrm{P}_0 + \theta\mathrm{P}_1$$

of the model has induced the very special property that all admissible procedures in Section 1.5 (Situation 2) are also 'Somewhere Minimum Risk'. Let us therefore now discuss a new example which is, at first sight, one step less regular.

A new example
Just as in the original example, a fair 'die' (now a regular simplex) has been rolled and Player I has to make a statement x about the true number of eyes y according to the following 'rules': when $y = 1$, Player I has to make statement $x = 1$, when $y = 2$ the statement must be $x = 2$, and if $y = 3$ or $y = 4$ Player I is allowed to choose between $x = 1$ and $x = 2$. Treatment of this example mimics that of the original one, though it looks a bit more complex. That, on the other hand, it is also more general than the original example will become clear during the analysis.

Now there are four possibilities for Player I to specify his choice of x, depending on the value y observed. These possibilities can be indicated as $\{(i,j);\ i,j \in \{0,1\}\}$ where $i = 1$ if Player I chooses $x = 1$ after seeing $y = 3$ (and $i = 2$ when $x = 2$ is chosen in this case) while $j = 1$ if Player I chooses $x = 1$ after seeing $y = 4$ (and $j = 0$ if he chooses $x = 2$). Our corresponding strategies will be indicated as $Q_{(i,j)}$ or Q_{2i+j}. Here Q_{2i+j} prescribes what we should infer if we assume that Player I chooses the possibility (i,j).

We can restrict, for trivial reasons, to procedures $Q_{a,b,c,d}$ as displayed in Table 1.7. Note that $Q_0 = Q_{0,0,1/3,1/3}$, $Q_1 = Q_{0,1/2,1/2,0}$, $Q_2 = Q_{1/2,0,0,1/2}$, and $Q_3 = Q_{1/3,1/3,0,0}$. Table 1.8, with probabilities $\mathrm{P}_{2i+j}(X = \xi, Y = \eta)$ is the analogue of Table 1.2.

The risks involved in the four possibilities (pure strategies (i,j) or $2i+j$ of Player I)

Table 1.7: Procedures $Q_{a,b,c,d}$

$q_\xi(\eta)$	$\eta = 1$	$\eta = 2$	$\eta = 3$	$\eta = 4$
$\xi = 1$	$1-a-b$	0	a	b
$\xi = 2$	0	$1-c-d$	c	d

$\mathrm{P}_{2i+j}(X=\xi, Y=\eta)$	$\eta=1$	$\eta=2$	$\eta=3$	$\eta=4$
$\xi=1$	$\frac{1}{4}$	0	$\frac{i}{4}$	$\frac{j}{4}$
$\xi=2$	0	$\frac{1}{4}$	$\frac{1-i}{4}$	$\frac{1-j}{4}$

Table 1.8: Procedures P_{2i+j}

are

$$\begin{aligned} R(0, Q_{a,b,c,d}) &= -\tfrac{1}{4}\log\left((1-a-b)(1-c-d)cd\right) \\ R(1, Q_{a,b,c,d}) &= -\tfrac{1}{4}\log\left((1-a-b)(1-c-d)cb\right) \\ R(2, Q_{a,b,c,d}) &= -\tfrac{1}{4}\log\left((1-a-b)(1-c-d)ad\right) \\ R(3, Q_{a,b,c,d}) &= -\tfrac{1}{4}\log\left((1-a-b)(1-c-d)ab\right), \end{aligned}$$

the envelope risks $R^*(k)$ are $R^*(0) = R^*(3) = \frac{3}{4}\log 3$ and $R^*(1) = R^*(2) = \log 2$ and the corresponding regrets (shortcomings)

$$\begin{aligned} S(0; Q_{a,b,c,d}) &= -\tfrac{1}{4}\log\left(3^3(1-a-b)(1-c-d)cd\right) \\ S(1; Q_{a,b,c,d}) &= -\tfrac{1}{4}\log\left(2^4(1-a-b)(1-c-d)cb\right) \\ S(2; Q_{a,b,c,d}) &= -\tfrac{1}{4}\log\left(2^4(1-a-b)(1-c-d)ad\right) \\ S(3; Q_{a,b,c,d}) &= -\tfrac{1}{4}\log\left(3^3(1-a-b)(1-c-d)ab\right). \end{aligned}$$

To construct the class $\mathcal{B} = \{Q_\rho \mid \rho \in S_3\}$ of all Bayes procedures, let ρ_1, ρ_2, and ρ_3 be a priori probabilities for possibilities 1, 2, and 3 (and hence $1-\rho_1-\rho_2-\rho_3$ for possibility 0). (It is trivial that all three $\rho_i \geq 0$ and that $\rho_1+\rho_2+\rho_3 \leq 1$.) The class of Bayes procedures consists (exactly) of the procedures that minimize the Bayes risks

$$(1-\rho_1-\rho_2-\rho_3)S(0) + \rho_1 S(1) + \rho_2 S(2) + \rho_3 S(3).$$

Elementary calculus provides the solutions $Q_{a,b,c,d}$ with

$$\begin{aligned} a &= \tfrac{\rho_2+\rho_3}{1+\rho_1+\rho_2+2\rho_3} \\ b &= \tfrac{\rho_1+\rho_3}{1+\rho_1+\rho_2+2\rho_3} \\ c &= \tfrac{1-(\rho_2+\rho_3)}{3-(\rho_1+\rho_2+2\rho_3)} \\ d &= \tfrac{1-(\rho_1+\rho_3)}{3-(\rho_1+\rho_2+2\rho_3)}. \end{aligned}$$

Here we used the average regret as Bayes risk. We could equally well have used the average risk, this would, obviously, have provided the same Bayes rule. The issues to be considered next are to define the *minimax risk* and *minimax regret* procedure.

The symmetry of $R(\cdot, Q_{a,b,c,d})$ suggests that the maximum risk is minimized if we can equalize an Bayes rule. This is obviously achieved if

$$\rho_1 = \rho_2 = \rho_3 = \tfrac{1}{4},$$

because the corresponding Bayes rule is such that

$$a = b = c = d = \tfrac{1}{4}.$$

Quite interesting, the minimax regret procedure is the same $Q_{1/4,1/4,1/4,1/4}$, but now with $\rho_1 = \rho_2 = \frac{1}{2}$ and,hence, $\rho_3 = 0$. If we plug in the above representations in ρ_1, ρ_2, ρ_3 for a, b, c, and d in the $S(\cdot, Q_{a,b,c,d})$, we obtain after observing symmetry in ρ_1 and ρ_2 (and hence $\rho_1 = \rho_2$) and $\rho_0 = 1 - \rho_1 - \rho_2 - \rho_3$ and ρ_3 (hence $\rho_0 = \rho_3$) that

$$(1 - \rho_1 - \rho_2 - \rho_3)S(0) + \rho_1 S(1) + \rho_2 S(2) + \rho_3 S(3)$$
$$= -\tfrac{1}{4}\left(2\rho_1 \log 2^4 + 2(\tfrac{1}{2} - \rho_1) \log 3^3 - 6 \log 2\right)$$

and, again,

$$a = b = c = d = \tfrac{1}{4}.$$

This coincidence of the minimax risk and minimax regret rule and the symmetry in the solutions shows that this second example is also not sufficiently general to be called paradigmic. Let us therefore consider a completely new example, in a different context.

1.8 An classroom example with Bernoulli trials

This chapter, so far, illustrates that straightforward conditioning to incorporate empirical evidence can be misleading. This phenomenon is well-known from other problems, such as the quiz-master paradox and the prisoner's dilemma (see Section 1.9 for formulations). The source of the information should be formalized and made part of the probabilistic model, which will then become 'statistical' in the sense that the unknown true value of a parameter appears. Hemelrijk (1978) states that 'the question of how to choose a statistical mathematical model has led to considerable confusion and controversy, and still does. Mathematical statisticians wisely save their skins by using the axiomatic approach, leaving the controversy to others and the confusion to the users of their theory. For axioms, however useful, say nothing about their application.'

Similar issues are involved elsewhere though they often go unnoticed. The following problem provides an example because neither information is provided about the player's set of alternatives nor about the rules he has to obey: 'A bridge player announces that his hand (of 13 cards) contains (i) an ace (that is, at least 1 ace), (ii) the ace of hearts. What is the probability that it will contain another ace?' (Parzen, 1960, p. 75).

From the statistician's viewpoint as well as from that of the epistemologist, this problem is 'ill-posed'. There, indeed, is a difference between the attitudes of the analyst (mathematician, probabilist) who proceeds on the basis of an elegant rationalization, and that of the statistician who has to start from the observation. In mathematical statistics we try to combine these rationalist and empiricist perspectives. In Sections 1.1 up to 1.6 we had considerable difficulty in obtaining consensus. In Section 1.7 the symmetry of the situation enforced some type of consensus. We shall now discuss a different elementary example, where, in the end, (approximate) consensus will be reached.

Problem
$X_1, X_2, \ldots$ are outcomes of independent Bernoulli trials, each with unknown success-rate p. The theoretical possibilities θ for p constitute the parameter space $\Theta = [0, 1]$. The outcome $x = (x_1, \ldots, x_n)$ observed has to be used to make an inference about the outcome y of $Y = X_n X_{n+1}$.

Exploration
Note that $y \in \mathcal{Y} = \{0, 1\}$. The 'inference' about y required can have the form of a distributional inference

$$Q(x) = (1 - \alpha(x))\epsilon_0 + \alpha(x)\epsilon_1.$$

Amongst other proper loss functions, the logarithmic loss function $L_{\log}(y, Q) = -\log q_x(y)$ and the Epstein scoring rule

$$L_{\mathrm{E}}(y, Q(x)) = \big(\mathbf{1}_{\{0\}}(y) - 1 + p\mathbf{1}_{\{1\}}(x)\big)^2 = \tfrac{1}{2} L_{\mathrm{B}}(y, Q(x))$$

can be used. Note that, in this case, the relationship between Epstein and Brier loss is such that $L_{\mathrm{E}}(y, Q(x)) = \frac{1}{2} L_{\mathrm{B}}(y, Q(x))$. The 'inference' $(x_n, y) = (0, 1)$ yields maximum loss (infinite loss if the logarithmic loss function is used) for all proper scoring rules, and $(x_n, y) = (0, 0)$ yields zero loss, because inferring $y = 0$ is obviously the only right choice after observing $x_n = 0$.

The inference about y required can also be identified as a point estimate (prediction) of y. Squared error loss is then attractive because $\mathbf{E}\,(Y - c)^2$ is minimum if

$$\begin{aligned} d(x_1, \ldots, x_n) &= \mathbf{E}\,(X_n X_{n+1} | X_1 = x_1, \ldots, X_n = x_n) \\ &= x_n \mathbf{E}\,(X_{n+1} | X_1 = x_1, \ldots, X_n = x_n) \\ &= x_n \mathbf{E}\,(X_{n+1}) \\ &= p x_n. \end{aligned}$$

Unfortunately, we do not know the true value p of the parameter θ. It is intuitively clear that the ideal 'solution' $d(x_1, \ldots, x_n) = p x_n$ should be replaced by a practical one where p is replaced by some estimate. The theoretician, however, might argue that $\mathbf{E}\,Y = p^2$ and that, hence, it is reasonable to use the best unbiased estimator[2]

$$d_1(x) = \frac{s_n(s_n-1)}{n(n-1)}$$

of p^2. Here and elsewhere $s_m = \sum_{i=0}^{m} x_i$ is used for notation. After some computations it is clear that the prediction risk is given by $\mathbf{E}\,(d_1(X_1,\ldots,X_n) - Y)^2 =$

$$\begin{aligned} &= (1-p)\mathbf{E}\,((d_1-Y)^2|X_n=0) + p\mathbf{E}\,((d_1-Y)^2|X_n=1) \\ &= \frac{(1-p)p^2(4p-2n(5p-2)-n^3(-2-p+p^2)+n^2(p^2+3p-4))}{(n-1)n^2}. \end{aligned}$$

An obvious drawback of d_1 is that it can happen that $d_1(x) > 0$ whilst $x_n = 0$ and therefore $Y = 0$ is certain. It seems more reasonable to start from the procedure $d_p(x) = px_n$ which would be used if p were known, and to replace p by an estimate, e.g.,

$$\begin{aligned} \hat{p} &= s_n/n \\ \hat{p} &= s_{n-1}/(n-1) \\ \hat{p} &= (s_n+1)/(n+2) && \text{(Laplace)} \\ \hat{p} &= (s_n+\sqrt{n}/2)/(n+\sqrt{n}) && \text{(Bernstein)} \end{aligned}$$

or, more generally,

$$\hat{p} = as_{n-1} + bx_n + c$$

where a, b and c have to be chosen such that

$$d_{a,b,c}(x) = x_n(as_{n-1} + bx_n + c)$$

displays good behaviour. Since $d_{a,b,c}(x) = ax_ns_{n-1} + (b+c)x_n$, it is obvious that we can generalize even more, by focussing on

$$d_{a,b}(x) = x_n(as_{n-1} + b).$$

Let us now for a moment look at the special case where $a = 1/(n-1)$ and $b = 0$. Here the actual risk is equal to $\mathbf{E}\;X_n^2((n-1)^{-1}S_{n-1} - X_{n+1})^2 = p^2(1-p)\frac{n}{n-1}$ and, hence, the risk function and regret function are

$$\begin{aligned} R(\theta, d_{1/(n-1),0}(x)) &= \frac{n}{n-1}\theta^2(1-\theta) \\ S(\theta, d_{1/(n-1),0}(x)) &= \frac{1}{n-1}\theta^2(1-\theta) \end{aligned}$$

[2] To construct the estimator d_1 start from X_1X_2. Because of independence, it is trivial that this is an unbiased estimator. Then use the fact that S_n is a sufficient statistic, and use Rao-Blackwell and Lehmann-Scheffé to find $d_1(x) = \mathbf{E}\,(X_1X_2|S_n)$.

and the latter one converges nicely to 0.

Bayesian methods
Another important special case is that where we try to minimize some integral

$$\int_0^1 R(\theta, d(x))\omega(\theta)\,\mathrm{d}\theta$$

by using Bayesian methods. If we use a Beta weight function, then the computations become feasible by invoking a Bayesian context with a probability distribution Q determined by $T \sim \text{Beta}(f, g)$ and $\tilde{X}_1, \ldots, \tilde{X}_{n+1} | (T = \theta)$ i.i.d. Bernoulli(θ). Interprete

$$\int_0^1 R(\theta, d) \frac{\theta^{f-1}(1-\theta)^{g-1}}{\beta(f, g)}\,\mathrm{d}\theta = \mathbf{E}\,\mathbf{E}\,(d(\tilde{X}_1, \ldots, \tilde{X}_n) - \tilde{X}_n \tilde{X}_{n+1})^2 | T,$$

this can be treated as usual by conditioning with respect to the observation, i.e. by writing the total expectation as

$$\mathbf{E}\,\mathbf{E}\,((d(\tilde{X}_1, \ldots, \tilde{X}_n) - \tilde{X}_n \tilde{X}_{n+1})^2 | \tilde{X}_1, \ldots, X_n)$$

and choosing d such that

$$\mathbf{E}\,((d(\tilde{x}_1, \ldots, \tilde{x}_n) - \tilde{x}_n \tilde{X}_{n+1})^2 | \tilde{X}_1, \ldots, X_n = x_1, \ldots, x_n)$$

is minimum, i.e.

$$\begin{aligned}
d(x) &= x_n \mathbf{E}\,(\tilde{X}_{n+1} | \tilde{X}_1, \ldots, X_n = x_1, \ldots, x_n) \\
&= x_n \mathrm{P}_Q(\tilde{X}_{n+1} = 1 | \tilde{X}_1, \ldots, X_n = x_1, \ldots, x_n) \\
&= x_n \frac{\mathrm{P}_Q(\tilde{X}_1 = \tilde{x}_1, \ldots, \tilde{X}_{n+1} = \tilde{x}_{n+1})}{\mathrm{P}_Q(\tilde{X}_1 = \tilde{x}_1, \ldots, \tilde{X}_n = \tilde{x}_n)} \\
&= x_n \frac{\beta(s_n + f + 1, n + g - s_n)}{\beta(s_n + f, n + g + 1 - s_n) + \beta(s_n + f + 1, n + g - s_n)} \\
&= x_n \frac{s_n + f}{n + f + g}
\end{aligned}$$

This corresponds to $d_{a,b}(x)$ when (a, b) relate to (f, g) such that $a = 1/(n + f + g)$, $b = (f + 1)/(n + f + g)$. Which f and g to choose remains unanswered here. Many other weight functions than Beta weight functions are possible, of course. Let us continue with the risk and regret function of $d_{a,b}$ We have

$$\begin{aligned}
\mathbf{E}\,(d_{a,b}(X) - X_n X_{n+1})^2 &= \mathbf{E}\,\mathbf{E}\,X_n^2 (aS_{n-1} + b - X_{n+1})^2 | X_n \\
&= p\mathbf{E}\,(aS_{n+1} + b - X_{n+1})^2 \\
&= p\left[a^2 \mathrm{Var} S_{n-1} + \mathrm{Var} X_{n+1} + ((an - a - 1)p + b)^2\right] \\
&= (a^2 n - a^2 + 1)p^2(1 - p) + p((an - a - 1)p + b)^2
\end{aligned}$$

and obtain the regret function

$$\begin{aligned} S(\theta, d_{a,b}) &= R(\theta, d_{a,b}) - \theta^2(1-\theta) \\ &= (n-1)a^2\theta^2(1-\theta) + \theta((an-a-1)\theta+b)^2 \end{aligned}$$

which we can study from various perspectives. One could try to find a class $\mathcal{D} \in \{d_{a,b}\}$ that minimizes the maximum regret. But also the classes of minimax risk rules, Bayes rules, etcetera, are of interest. In practice, the induced inferences will be approximately equal.

1.9 Related well-known examples

The following problems are related to the die-rolling games of this chapter.

The quiz-master problem
Also known as the three-doors problem and the Monty Hall dilemma, this is one of the most well-known paradoxical problems in popular statistics. The problem is as follows: 'Suppose you are a contestant in a TV game show and you made it to the grand final. This is your chance to win a brand new car. The game show host shows you three doors, behind one of them is the car, behind the other two goats. To take the car home, you have to pick the right door. After you have picked a door, say door A, the host (who knows what's behind which door) opens one of the other doors, say B, showing you a goat. He gives you the opportunity to switch to door C. Is this to your advantage?' Kooi (1999)

Many people's initial intuition will say 'No' to this question. You picked at random, all probabilities must be $\frac{1}{3}$. However, these of course 'change' when door B is opened. A simple probabilistic exploration of the problem shows that it is in your advantage to change doors, since the car is behind door C with probability 2/3. This only holds in the context where both game show host and contestant are 'machines' who will choose at random between the available options, with equal probabilities.

This problem is known for about 40 years, see e.g. Mosteller (1965, p. 4) and Selvin (1975). Another reference is Kooi (1999), where the problem is dealt with using Probabilistic Epistemic Logic.

The prisoner's dilemma
This problem is probably even older then the quiz-master problem (according to Rapoport (1974) the first appearance of this problem was in 1952 by Flood). One version of this two-player nonconstant-sum game is as follows: 'two prisoners who escaped and participated in a robbery have been recaptured and are awaiting trial for their new crime. Although both are guilty, the district attorney is not sure he has enough evidence for conviction. To entice them to testify against each other, the D.A. tells each prisoner the following:

> ' "If only one of you confesses and testifies against your partner, the person who confesses will go free while the person who does not confess will surely

> be convicted and given a 20-year jail sentence. If both of you confess, you will both be convicted and sent to prison for 5 years. If neither of you confesses, I can convict you both of a misdemeanor and you will each get 1 year in prison." What should each prisoner do?'

(Winston, 1994, p. 850). On this dilemma, entire books are written. In a mathematical-statistical context, the ones by Rapoport et al. (1965) and Poundstone (1992) are very interesting.

Again, the solution depends heavily on the information of the game being the knowledge the two prisoners have about each other. An advantage of our problem is that it is less vexatious and more in line with what we have to do in mathematical statistics at large and in distributional inference in particular.

The three prisoners problem

> 'Three prisoners, A, B, and C are locked in their cells. It is common knowledge that one of them will be executed the next day and the others will be pardoned. Only the Governor knows which one will be executed. Prisoner A asks the guard a favor: "Please ask the governor who will be executed, and then take a message to one of my friends B and C to let him know that he will be pardoned in the morning." The guard agrees, and comes back later and tells A that he gave the pardon message to B. What are A's changes of being executed, given this information?'

(Russel and Norvig, 1995, ex. 14.12). See also Diaconis and Zabell (1986). This problem attracted our attention through Op den Akker (1998).

Similar to the quiz-master paradox, inferences about the probability (that A is executed) depend on the information-strategy, this time of the guard. Quite interestingly, this is not remarked in neither Russel and Norvig (1995) nor Op den Akker (1998).

Finally we want to mention yet another example, namely the two-envelope paradox. This paradox, and its accompanying problem, will constitute the core of the following chapter. There, the problem will be dealt with in a logical, probabilistic, statistical and game-theoretic way.